PG-DSBA Dec_A 2020

Advanced Statistics Project

Submitted by: F Maria Jasmine

Date of Submission: 19/03/2021

Table of Contents

Problem 1A:

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.

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

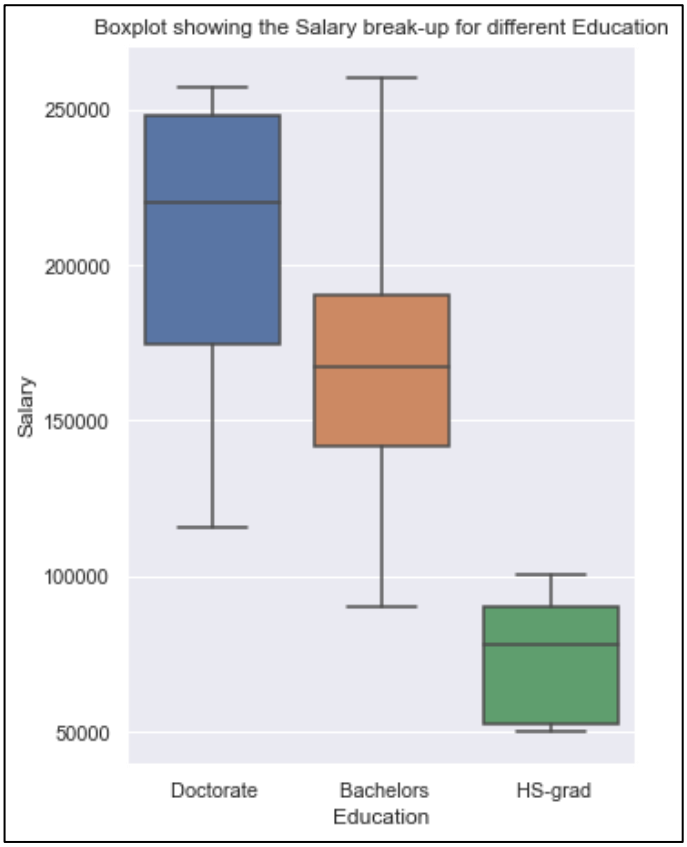
Data Description:

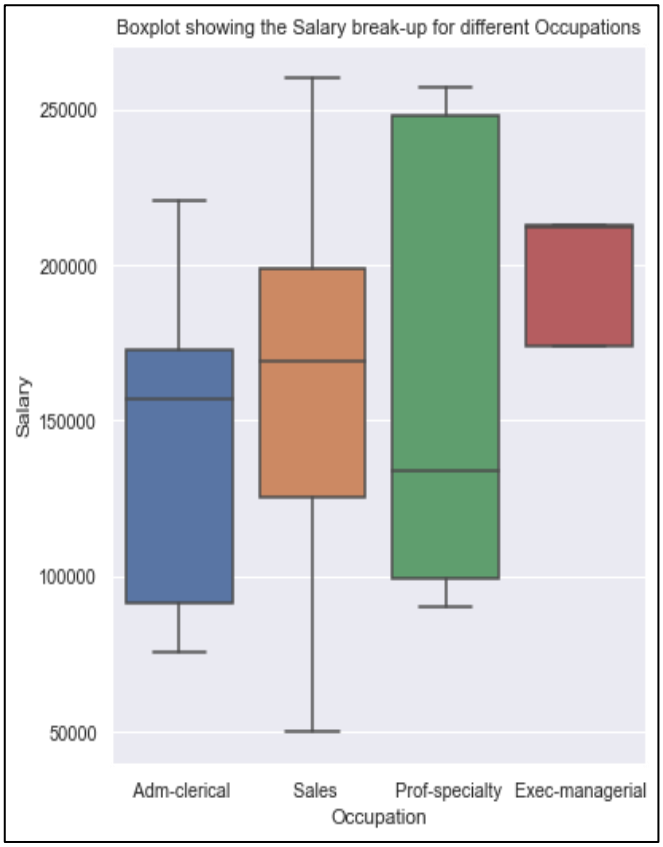
1. The Salary Data has 40 rows and 3 columns
2. There are 2 categorical and 1 integer data type
3. Below is the summary of the target variable Salary:

	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

4. There are no anomalies, missing values and duplicates in the data.

Bivariate Analysis:





- No outliers in the variable Salary.

Salary Breakup for each combination of Education and Occupation:

Row Labels	Admin-clerical	Exec-managerial	Prof-specialty	Sales	Grand Total
Bachelors	512133	772807	423151	769203	2477294
Doctorate	665831	212781	1486637	969583	3334832
HS-grad	236279	0	286603	152467	675349
Grand Total	1414243	985588	2196391	1891253	6487475

- Salaries of Individuals who have Doctorate is higher than other Education levels.
- Salaries of Individuals who are Professionals or Specialist is higher than other Occupation types.

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

Hypothesis for One-Way Anova on salary with respect to Education:

H_0: $\mu 1 = \mu 2 = \mu 3$ i.e., The mean salary for different Education levels is equal

Education is not a significant cause for Salary to either increase or decrease.

H_1: At least for one Education level the mean salary is different from the rest

Education is a significant cause for Salary to either increase or decrease.

Hypothesis for One-Way Anova on salary with respect to Occupation:

H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$ i.e., The mean salary for different Occupation levels is equal

Occupation is not a significant cause for Salary to either increase or decrease.

H_1: At least for one Occupation type the salary is different from the rest

Occupation is a significant cause for Salary to either increase or decrease

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.

One-Way Anova:

1. One-way ANOVA is a hypothesis test, used to test the equality of three or more population means simultaneously using variance.
2. It should have One independent variable.
3. Three or more levels for one factor

Assumptions of One-Way Anova:

1. The samples drawn from different populations are independent and random.
2. The response variables are normally distributed.
3. The variances of the population means are equal.

Mean Salary for the factor Education at different levels:

Below is the table showing the break-up of means of different Education levels with respect to Salary:

Row Labels	Means
Bachelors	165152.93
Doctorate	208427.00
HS-grad	75038.78

The mean for HS-grad is significantly low when compared with other levels. By comparing the means, we find that, there is difference between the means of Education level with respect to Salary. One-way Anova is performed to prove or disprove the above interpretation.

Hypothesis for One-Way Anova on salary with respect to Education:

H_0: $\mu 1 = \mu 2 = \mu 3$ i.e., The mean salary for different Education levels is equal

H_1: At least for one Education level the mean salary is different from the rest

One-Way ANOVA Table:

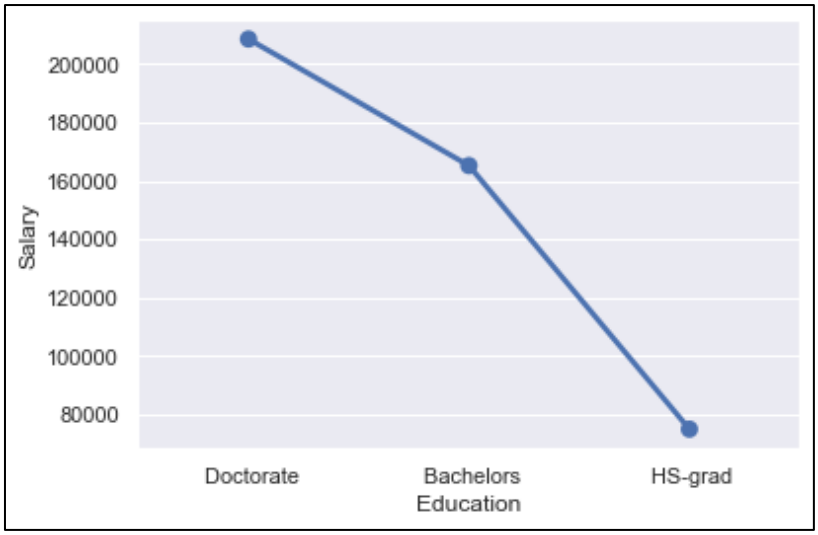
ANOVA table on Salary with respect to Education

	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

Interpretation of ANOVA table:

1. The degree of freedom: Education level has 3 levels, So the df corresponding to education is (3-1=2) and df for residual is (40-3=37)
2. The F-statistic is 30.9 and F-Critical is 3.25. The F-stat falls within the critical region. We reject the Null hypothesis.
3. The p-value is significantly less than the level of significance (5%). We reject the null hypothesis of equality of means.
4. The difference between the variables is 30.9 times larger than the difference within the variables.

Graphical representation: Point plot:



Conclusions:

1. There is a significant cause effect relationship between Education and Salary.
2. The salary depends on Educational level.
3. Therefore, at 95% level of Confidence, we can conclude that the Educational level is the significant cause for the Salary to either increase or decrease.

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

Mean Salary for the factor Occupation at different levels:

Below is the table showing the mean salary break-up for 4 different types of the Occupation.

Row Labels	Means
Admin-clerical	141424.30
Exec-managerial	197117.60
Prof-specialty	168953.15
Sales	157604.42

The variations of mean salary among all the four different Occupation types are not highly differentiated. One-Way Anova is performed to check if the means for all occupation levels are equal or different.

Hypothesis for One-Way Anova on salary with respect to Occupation:

H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$ i.e., The mean salary for different Occupation levels is equal

H_1: At least for one Occupation type the salary is different from the rest

One-Way ANOVA Table:

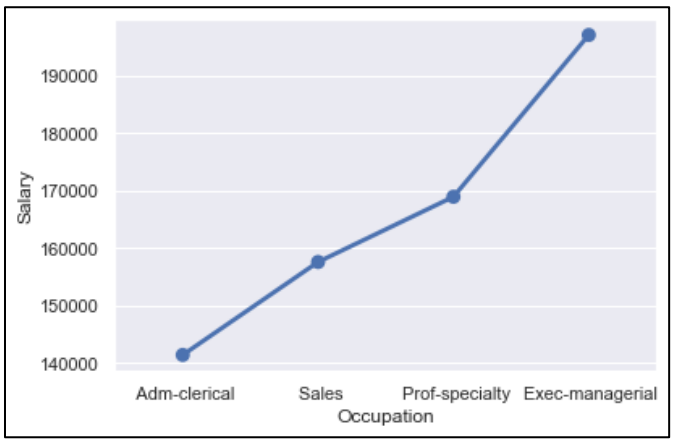
ANOVA table on Salary with respect to Occupation

	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

Interpretation of Anova Table:

1. The degree of freedom: The factor occupation has 4 levels, the df corresponding to occupation is (4-1=3) and df for residual is (40-4=36)
2. The F-statistic is 0.88 and the F-Critical is 2.866. The F-stat does not fall within the critical region. So, we fail to reject the Null hypothesis
3. The P-value is more than the significance level (5%). The null hypothesis for equality of means is not rejected.
4. The difference between the variables is only 0.88 times larger than the difference within the variables.

Graphical representation: Point plot:



Conclusions:

1. There is no significant cause effect relationship between Occupation and Salary.
2. The changes in Occupation does not affect the response variable Salary.
3. Therefore, at 95% confidence level, we can conclude that Occupation is not a significant cause for the Salary to increase or decrease.

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.

Tukey Test for the variable Education:

Null hypothesis is rejected and alternate hypothesis is accepted. Here, there is significant difference in salary when there is a change in Education.

To find out which class/ level of Education is significantly different, Tukey test (Multi Comparison) is used.

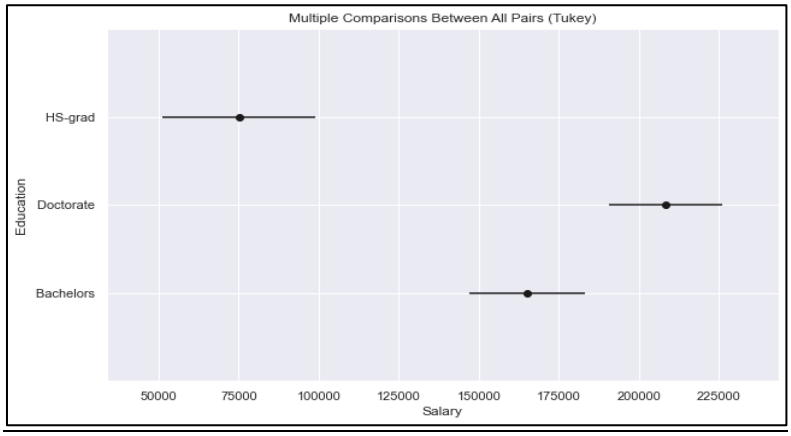
Multiple Comparison of Means - Tukey HSD, FWER=0.05

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

Interpretations:

1. p-values for all the groups are significantly lesser than the alpha.
2. For all the groups we reject the null hypothesis of equality of means.
3. The pairs Bachelors & HS-Grad and Doctorate & HS-Grad have very low p-value. Which means the common class among the two pairs HS-Grad is very far from the mean.
4. Hence, HS-Grad is the significantly different from the other two class.
5. The mean salary of HS-Grad is 75038 whereas the mean salary of the other two class is 165153 and 208427

Graphical representation of Tukey Test:



Tukey Test for the variable Occupation:

Null hypothesis of equality of means is not rejected for the variable Occupation. Here, there is significant no difference in salary when there is a change in Education. Let's run a tukey test to check the behaviour of the pairs.

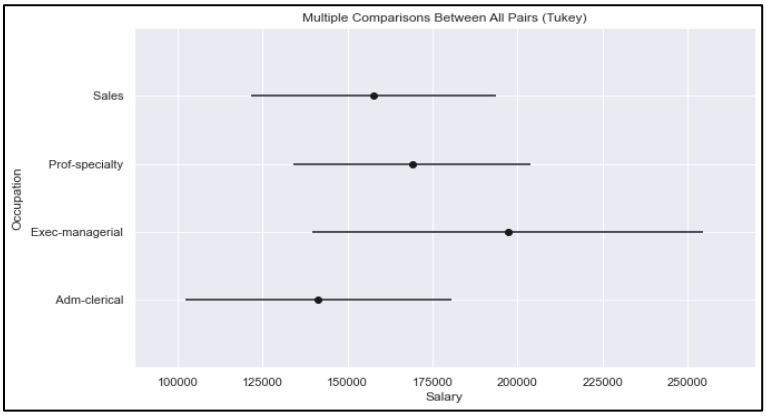
Multiple Comparison of Means - Tukey HSD, FWER=0.05

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

Interpretations:

1. The p-values for all the groups are higher than the alpha.
2. We fail to reject the null hypothesis of equality of means.
3. All the groups are almost near to the mean salary.
4. The classes like Admin clerical and Exec managerial is slightly far from the mean when compared with other groups.

Graphical representation of Tukey Test:

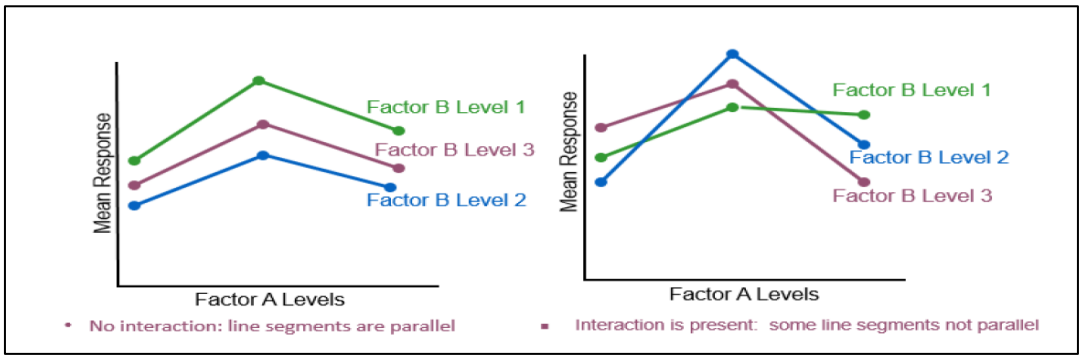


1.5 ***What is the interaction between the two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.***

What is interaction between the two treatments:

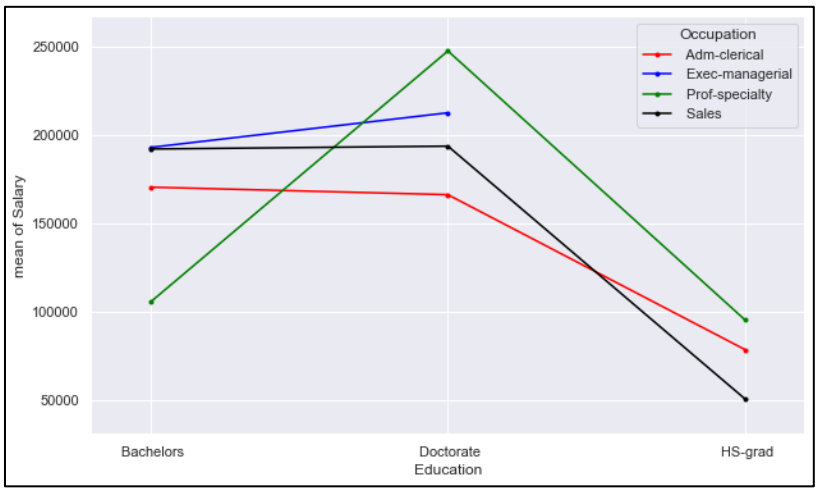
Interaction is a quantification of association of two factors. If one factor behaves differently at different levels of one or more factors, an interaction effect is said to exist. Interaction occurs when the pattern of the cell means in one row (going across columns) varies from the patterns of cell means in other rows.

Graphical representation of Interaction effect of one variable with other:



The above graph is the representation of Non-interaction and interaction between two factors.

For the given data set the interaction plot is mapped. Below is the pictorial representation of the interaction effects.



From the above plot, we can understand that there is significant interaction between the Variables Education and Occupation. To check the same, ANOVA is performed with interaction effects.

Hypothesis for Interaction effects of salary with respect to Occupation:

*H*0: There is no interaction between Education and Occupation with respect to Salary

*H*1: There is a significant interaction between Education and Occupation with respect to Salary

Interaction effects ANOVA Table:

```
ANOVA Interaction table on Salary with respect to Education and Occupation
------------------------------------------------------------------------
                         df        sum_sq       mean_sq          F  \
Education:Occupation  11.0  1.438019e+11  1.307290e+10  18.384842
Residual              29.0  2.062102e+10  7.110697e+08        NaN

                            PR(>F)
Education:Occupation  3.340466e-10
Residual                      NaN
```

Interpretation of ANOVA Table:

1. The degree of freedom: The interaction effect has 12 levels, so the corresponding df is (12-1=11) and df for residual is (40-11=29)
2. The F stat is 18.38 which falls within the F critical, so we reject the null hypothesis of no interaction effect.
3. The p-value is significantly less than the level of significance, so we reject the null hypothesis and prove that there exists an interaction effect.

Conclusions:

1. One way ANOVA of Education proved that the factor Education is the significant cause for the Salary to increase or decrease. However, the ANOVA conducted for Occupation confirmed that there is no significant dependency of salary with respect to Occupation.
2. When the two factors Education and Occupation are treated together, both Education and Occupation becomes the significant cause for the salary to either increase or decrease.
3. It is concluded that, there exists an interaction between the two factors which shows some impact on the Salary.

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?

Two Way ANOVA:

1. It is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
2. Two independent variables
3. Effects of multiple level of two factors.

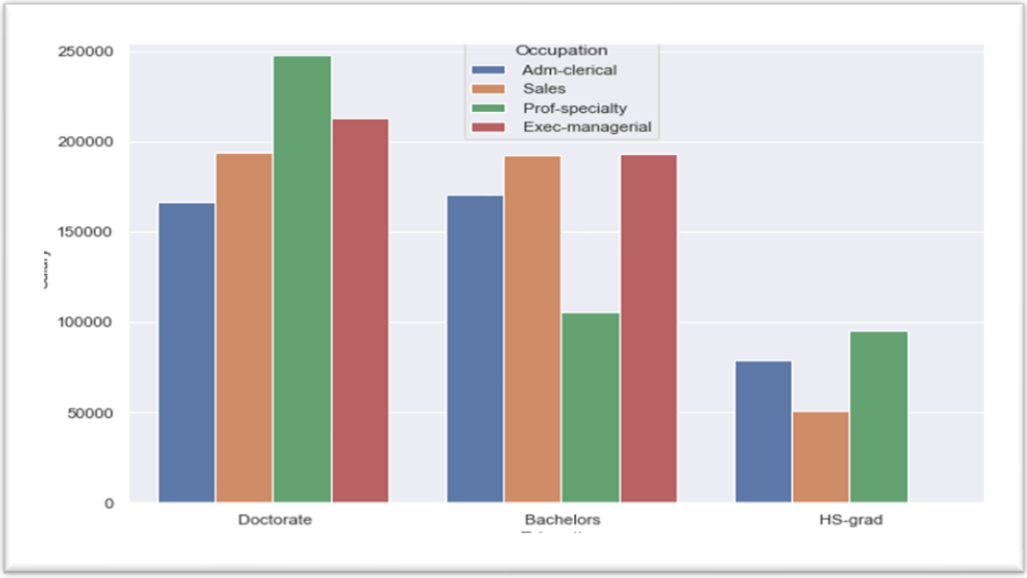
Assumptions for Two-Way ANOVA:

1. Dependent variable should be measured on a continuous scale.
2. The two independent variables should be categorical and independent groups.
3. Independent and random samples.
4. The response variables are normally distributed.
5. The variances of the population means are equal.

Mean Salary at each combination of Education and Occupation type:

Row Labels	Admin-clerical	Exec-managerial	Prof-specialty	Sales	Grand Total
Bachelors	170711.00	193201.75	105787.75	192300.75	165152.93
Doctorate	166457.75	212781.00	247772.83	193916.60	208427.00
HS-grad	78759.67	0.00	95534.33	50822.33	75038.78
Grand Total	**141424.30**	**197117.60**	**168953.15**	**157604.42**	**162186.88**

Graphical representation: Bar plot showing the Salary changes with respect to the two factos:



Two-Way Anova Table:

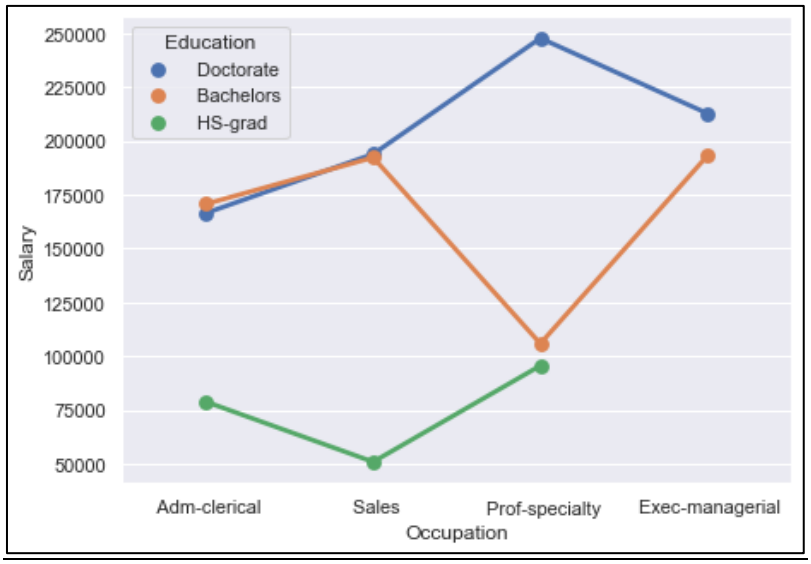
```
Two-Way Anova Table
------------------------------------------------------------------------
                        df        sum_sq       mean_sq          F  \
Education              2.0  1.026955e+11  5.134773e+10  72.211958
Occupation             3.0  5.519946e+09  1.839982e+09   2.587626
Education:Occupation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

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

Interpretations of ANOVA Table:

1. The degree of Freedom: For the factor Education and Occupation there are 3 and 4 levels, so the df corresponding to the both factors are 2 and 3 respectively.
2. For the interaction between the Education and Occupation, the df is (7-1=6)

3. The p-value for Education is significantly less than alpha. So, we reject the null hypothesis of equality of means. We conclude that there is a cause-and-effect relationship between Education and Salary.
4. The p-value for Education is not less than alpha. So, we fail to reject the null hypothesis of equality of means. We conclude that there is no cause-and-effect relationship between Occupation and Salary.
5. The p-value for the interaction effect is less than the alpha. So, the equality of means at each combination of Education and Occupation is also rejected.

Graphical representation: Point Plot



Conclusion:

1. As more factors are being included in the model, SSE is being reduced. The aim of ANOVA is to explain the total variability in the data, i.e., to assign the variability to definitive causes.
2. All three hypotheses are significant at 5% level. Therefore, our conclusion based on two-way ANOVA test, we reject the equality of means for the factor Education and we fail to reject the equality of means for Occupation.
3. Similarly, equality of means at each combination of Education and Occupation is also rejected, which means there is an interaction effect for the two treatments.

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

Business Implications:

1. The problem statement started with the sentence '*Salary is hypothesized to depend on educational qualification and occupation'*. After performing ANOVA, it is now clear that Salary depends on Education and there is no evidence for dependency of Salary with respect to Occupation.
2. The problem statement is not clear about the purpose of the data. So, let's consider two scenarios where the selection of correct factors gives the expected output.
3. Marketing Survey/Research for analysing the Target Market: In this scenario, the analyser should only consider the Education factor, as it has a significant cause-and-effect relationship with Salary. An increase or decrease in Salary totally depends on the factor Education. So, for categorising the customers on the basis of Education should give the different breakups of Salary among the target market as well.
4. Recruitment/HR view: To determine the Salary for an individual who wish to join a new firm, both the Education and previous Occupation has to considered. As there exists an interaction effect between both, it is wise to consider both factors in this scenario.

Conclusions:

It is observed that the variation in Salary is significantly impacted by Education and not by the Occupation. There is an impact in Salary with their interaction effect. ANOVA helps in identifying which independent factor(s) can explain the variation in the response variable.

Problem 2:

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

Data Description:

1. The given data set has 777 Rows and 18 Columns.
2. The data types are 1 object, 1 float and 16 integer type of columns.
3. There are no duplicates, missing values and anomalies in the given data.
4. The screenshot of the summary of the data is presented below:

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

Anomaly:

The graduation rate is more than 100% for one particular college. It is treated and imputed with 100%.

Screenshot of the Graduation rate description is attached below:

```
count    777.000000
mean      65.440154
std       17.118804
min       10.000000
25%       53.000000
50%       65.000000
75%       78.000000
max      100.000000
Name: Grad_Rate, dtype: float64
```

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

Exploratory Data Analysis:

1. Univariate Analysis: This provide us with descriptions of single variables we are interested in using in more advanced tests and help us narrow down exactly what types of bivariate and multivariate analyses we should carry out.
2. Multivariate analysis involves observation and analysis of more than one statistical outcome variable at a time

Univariate Analysis:

1. For the given data set Histogram and Boxplot from Seaborn is used for Univariate analysis.
2. Column names are renamed for easy and clear understanding. The new column names are listed below:

0	Names	0	Univ/College	10	Books	10	Cost_Books
1	Apps	1	Received_Applications	11	Personal	11	Personal_Exp
2	Accept	2	Accepted_Applications	12	PhD	12	Faculty_Phd
3	Enroll	3	Students_Enrolled	13	Terminal	13	Faculty_Terminal
4	Top10perc	4	Students_Top10	14	S.F.Ratio	14	SF_Ratio
5	Top25perc	5	Students_Top25	15	perc.alumni	15	Perc_alumni_donation
6	F.Undergrad	6	Full_time	16	Expend	16	Instruct_Exp
7	P.Undergrad	7	Part_time	17	Grad.Rate	17	Grad_Rate
8	Outstate	8	Out_of_State_Stud				
9	Room.Board	9	Cost_Room				

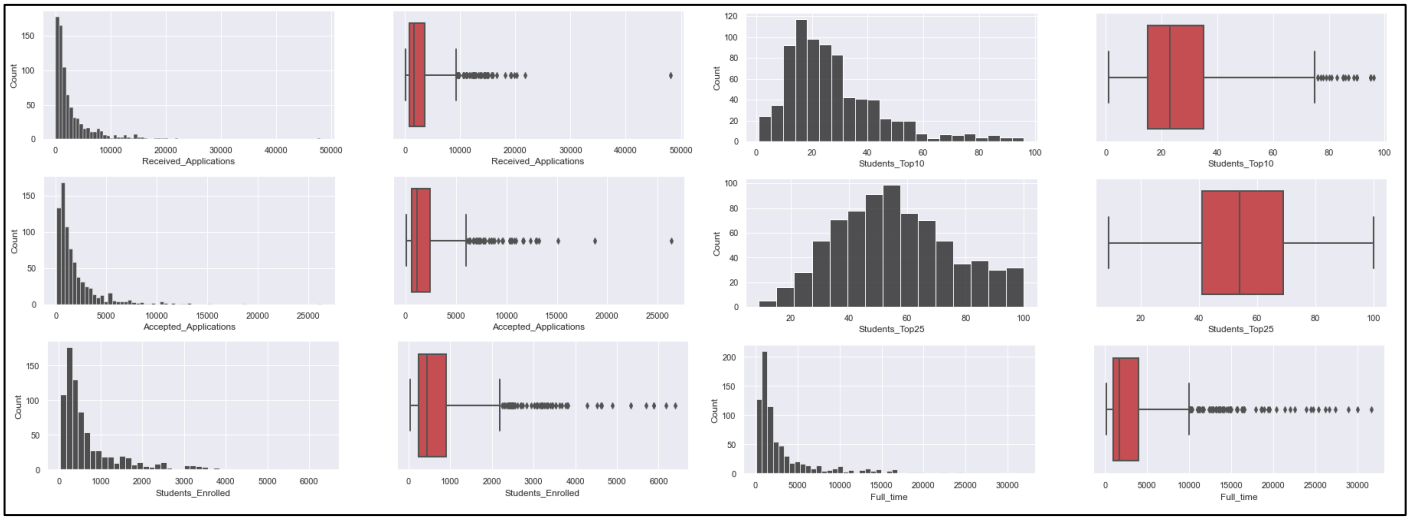
3. For performing univariate analysis, only the numeric columns are considered. The Univ/College column is removed.

Insights from Univariate Analysis:

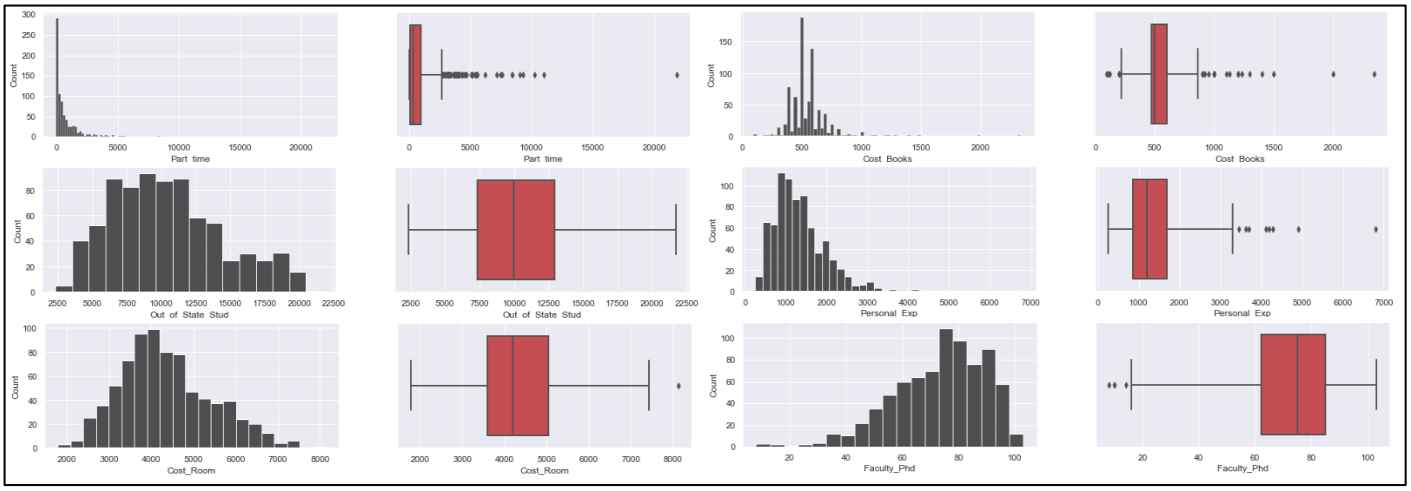
- On an average 67% of applications are accepted by the University/College
- Out of 67% accepted applications only an average of 25% students are enrolled.
- Column Student_Top25 is almost normally distributed and has no outliers.
- Students prefer Full-time graduation than Part-time.
- Cost of Room for Out state students range from 1780 to 8120. There is an adequate choice of selection of room with respect to cost.
- Almost 50% of the cost of books amounts to 500. Only 25% of the books are more than 600.
- Almost 75% of students spend less than 1700 for personal expenses.
- On an average 72% of Faculties have PhD and 79% have Terminal degrees
- A good Student to Faculty ratio would be 16/18 students each per faculty. Out data set shows an average of 14 students per faculty. Which means for every 14 students there is a faculty.
- Only 25% of colleges have an instructional expenditure above 10000 approximately.
- The low the graduation rate, the quality of the institution is at stake. 75% of the institutions have only 78% as graduation rate.

Graphical representation: Histogram and Boxplot for all the Numeric Columns:

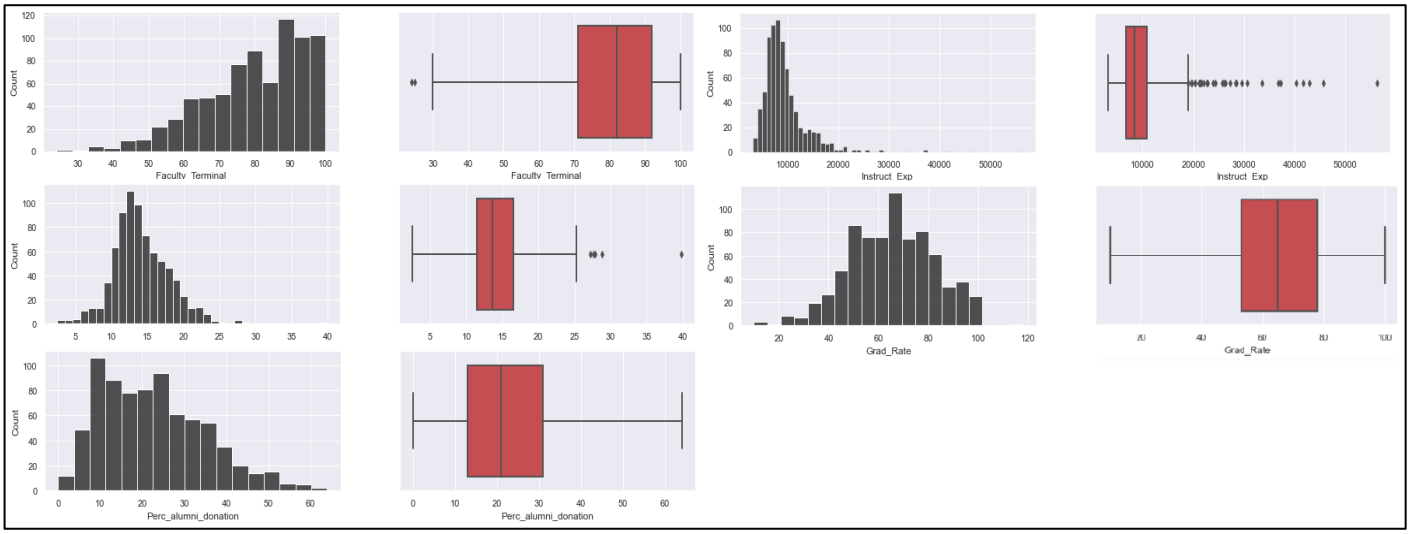
1.Received applications 2. Accepted applications 3. Students enrolled 4. Students Top10 5. Students Top 25 6. Full time



7.Part time 8. Out of state 9. Cost of room 10. Cost of books 11. Personal Exp 12. Faculty PhD

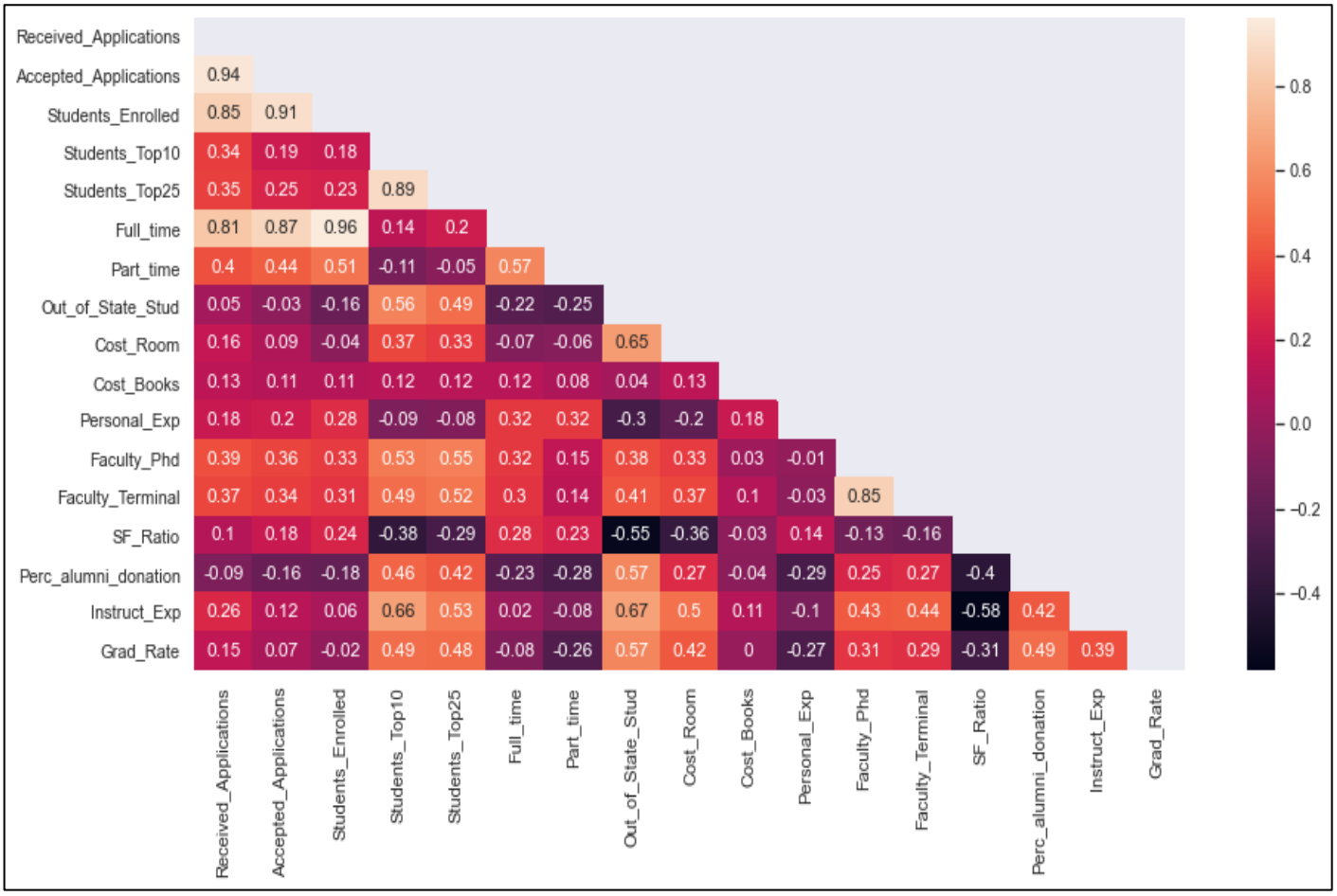


13. Faculty Terminal 14. SF Ratio 15. Percentage of alumni donation 16. Instructional Expenditure 17. Graduation Rate



Bivariate Analysis:

Heat map from Seaborn using the correlation matrix is performed to understand the relationship between any two variables.



Insights from Bivariate Analysis (Heatmap):

1. Negatively correlated Variables:

- Instruction expenditure includes expenditure for teachers' salaries and benefits for teachers. *Instructional expenditure and SF Ratio* shows a strong negative correlation. Which means, increase in expenditure does not increase the student faculty ratio.
- *Graduation rate and SF Ratio* show strong negative correlation. An increase in SF ratio alone does not boost the graduation rate.
- *Cost of room and students enrolled* is negatively correlated. Higher the cost of room lower is the students enrolled for both Full-time and part-time.
- *SF Ratio and Faculty PhD and Terminal degree* are negatively correlated. A faculty with PhD or Terminal degree can handle a greater number of students irrespective of the SF Ratio.

2. Positively correlated Variables:

- *Graduation rate, Faculty PhD, Faculty Terminal, Cost of room and Instructional expenditure* shows positive correlation. When more Faculties hold PhD and terminal degrees, the graduation rate tends to be high and students are willing to pay more instructional expenditures and cost for room.

- *Students enrolled and full time* is strongly correlated. Students highly prefer full time course.
- Students from Top 10 and 25 Higher secondary class look for a greater number of faculties with PhD and Terminal degree and does not consider SF ratio as a key element for selection of any institution.
- When percentage of alumni donation is high the cost of books are less. They are negatively correlated.
- Percentage of Alumni donation is very much correlated with Top10 and 25 students, faculties and Graduation rate.

Multi-variate Analysis:

For the purpose of multi variate analysis, *Scatterplot, Relational plot, Linear model plot and Pair plot are used.*

Scatter plots are the graphs that present the relationship between two variables in a data-set.

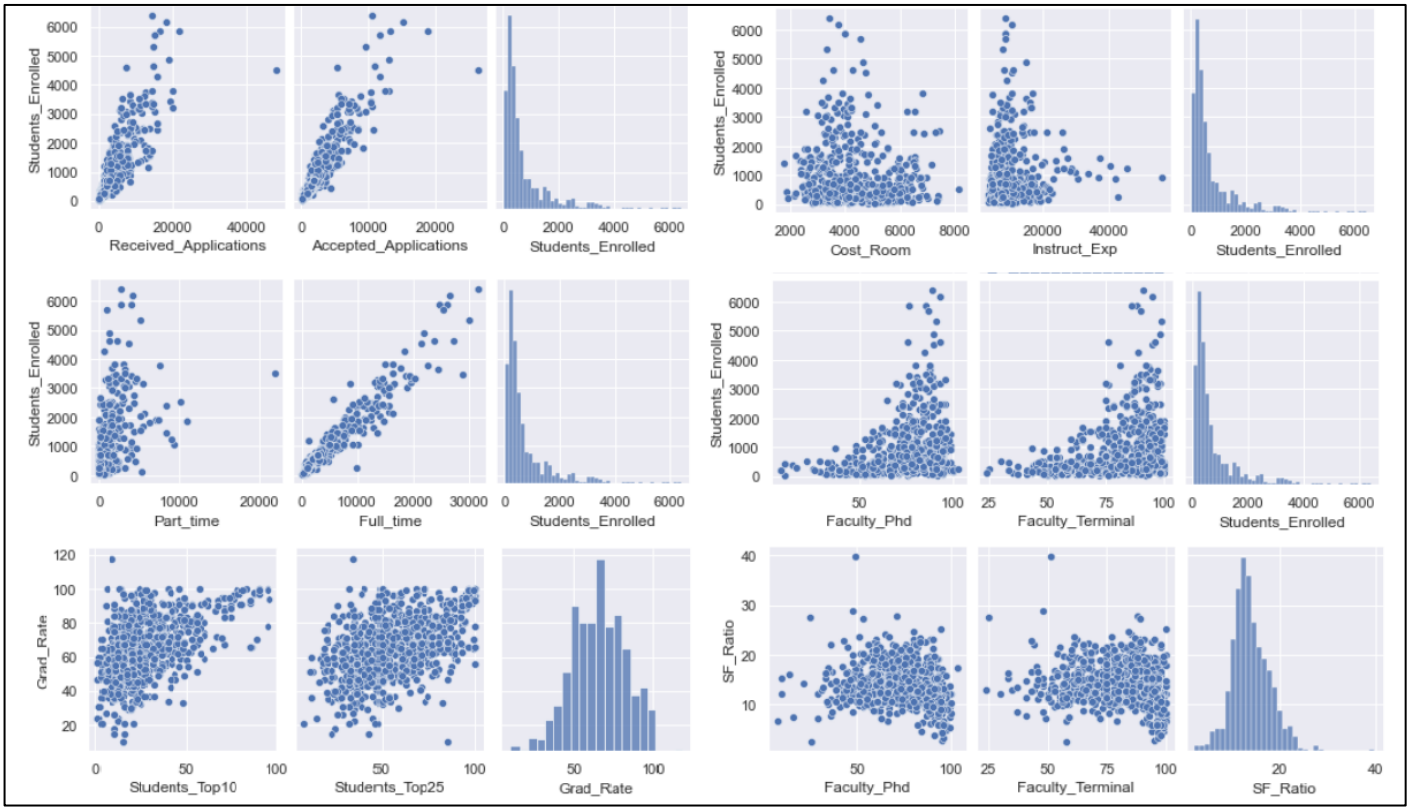
Linear model plot is used to get a best fit line to out plot

Pair plot allows us to see both distribution of single variables and relationships between two variables.

Pair Plots:

Below are the plots of different combinations of variables to understand their behaviour.

The variables that are analysed using the pair plots are: Students Enrolled, Full time students, Part time students, received applications, Accepted applications, Cost of room, Instructional expenses, Out of State students, SF Ratio, Faculty PhD, Faculty Terminal and SF Ratio.



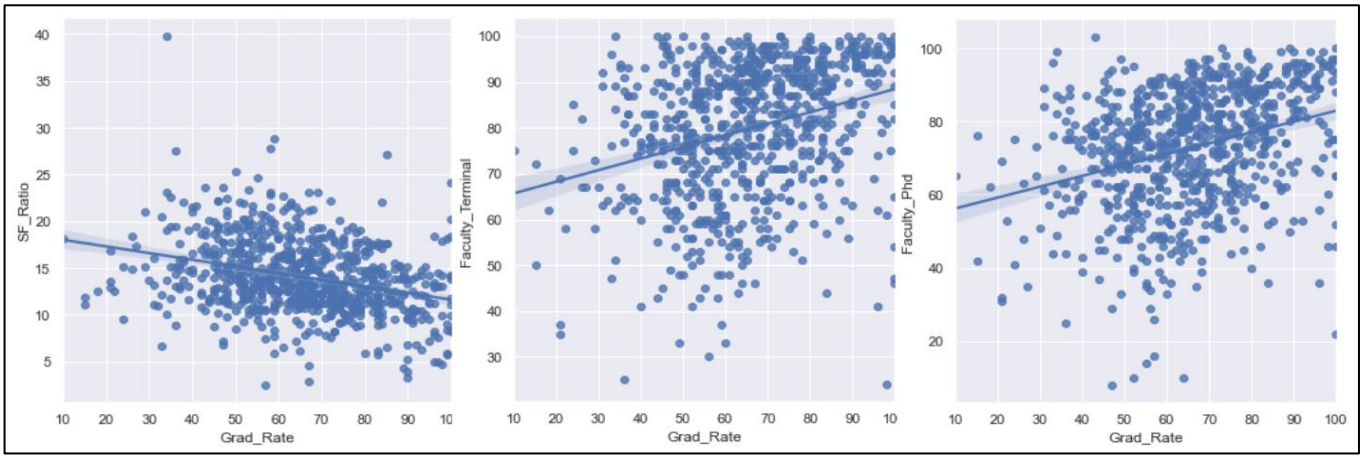
Insights from Pair plot:

- The more the applications received and accepted, the more is the students enrolled.
- Students prefer Full time over Part time graduation.
- Institutions with a greater number of Top10 and Top25 students from Higher Secondary class, has a high graduation rate.
- Students are comfortable to pay around 4000-4500 as room charges.
- The institutions with high number of faculties having a PhD and Terminal degree have more student's enrolment.
- SF Ratio tends to be lower for the institutions with high PhD and terminal degree faculties.

***Linear model (regression line) plot**:*

To understand the relationship between SF Ratio, Graduation Rate, Number of faculties with Terminal degree and instructional expenditure linear model plot is used.

Below is the graphical representation of the lm-plot



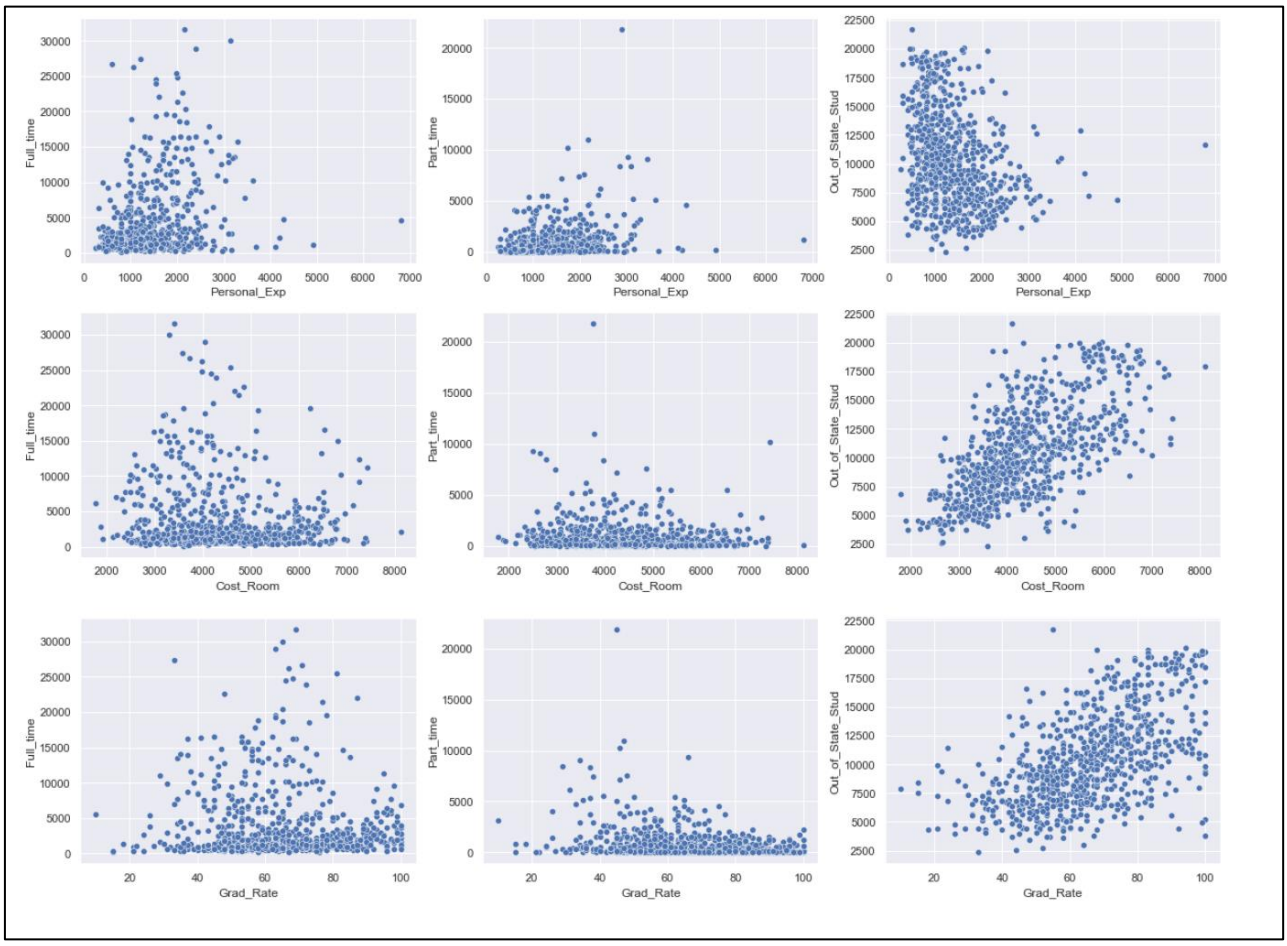
Insights from lm-plot:

- A good Student-to-Faculty ratio is 16 or 18 students per each faculty. The first plot clearly shows, *the graduation rate is high for the SF Ratio less than 20.*
- The graduation rate increases with an increase in number of Faculties gaining a Terminal degree.
- The instructional expenditure increases with an increase in Graduation rate for a particular institution.

Scatter-plot:

To understand the relationship between Full time, Part time and Out state students with Personal Expenses, Cost of room and Graduation Rate, scatter plot is used.

Below is the graphical representation of the scatter plot.



Insights from Scatter-plot:

- All Students from Out state have a need for spending for personal expenses than Full time and Part time students.
- The cost of room increases with the increase in count of out-state students.
- More number of out state students are graduated when compared with Full time and Part time.

Therefore, from EDA the key variables are identified. These variables can be considered an important criterion for any student to join an institution.

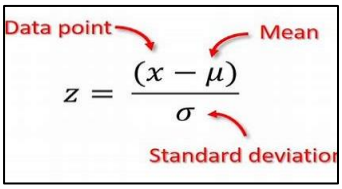
Number of Applications received and Students actually enrolled, Top 10 and Top 25 students, Cost of books, Cost of room, Alumni donation, SF Ratio, Faculty quality and Graduation rate

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

Need for Scaling:

- When dealing with data that has features with different scales, it's often important to scale the data first.
- In the given data set we have different scales like, Percentages, Ratio, Count and Cost. Hence, it is mandatory to scale data before performing PCA.
- The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
- More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (For example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.

Scaling by Z-score method:



- Mathematically this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable. Z score scaling method is imported from SciPy. Stats library
- Below is the screenshot of the Scaled data obtained.

	Received_Applications	Accepted_Applications	Students_Enrolled	Students_Top10	Students_Top25	Full_time	Part_time
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005

- The mean of the Scaled data becomes zero
- The Standard deviation becomes 1.
- Below is the screenshot of the description of the Scaled data, showing mean and the standard deviation.

	count	mean	std	min	25%	50%	75%	max
Received_Applications	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accepted_Applications	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Students_Enrolled	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Students_Top10	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Students_Top25	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391

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

Covariance matrix: Total variance of individual dimensions and across the dimensions

Correlation matrix: It is a table showing the correlation coefficients between the variables after Scaling the data.

Centering is crucial [(X – Xbar) where X bar becomes 0 (mean is 0 after scaling)] for the rotating the matrix.

The new dimension captures the maximum variance in the data points and also reduces the total error of the representation.

The covariance matrix which has become the correlation matrix after standardization, changes after rotating the axis.

In rotated axis, all the variables are captured by the Principal components themselves and the off-diagonal elements are no longer correlated.

Covariance Matrix:

	Received_Applications	Accepted_Applications	Students_Enrolled	Students_Top10	Students_Top25	Full_time	Part_time
Received_Applications	1.00	0.94	0.85	0.34	0.35	0.82	0.40
Accepted_Applications	0.94	1.00	0.91	0.19	0.25	0.88	0.44
Students_Enrolled	0.85	0.91	1.00	0.18	0.23	0.97	0.51
Students_Top10	0.34	0.19	0.18	1.00	0.89	0.14	-0.11
Students_Top25	0.35	0.25	0.23	0.89	1.00	0.20	-0.05

Correlation Matrix:

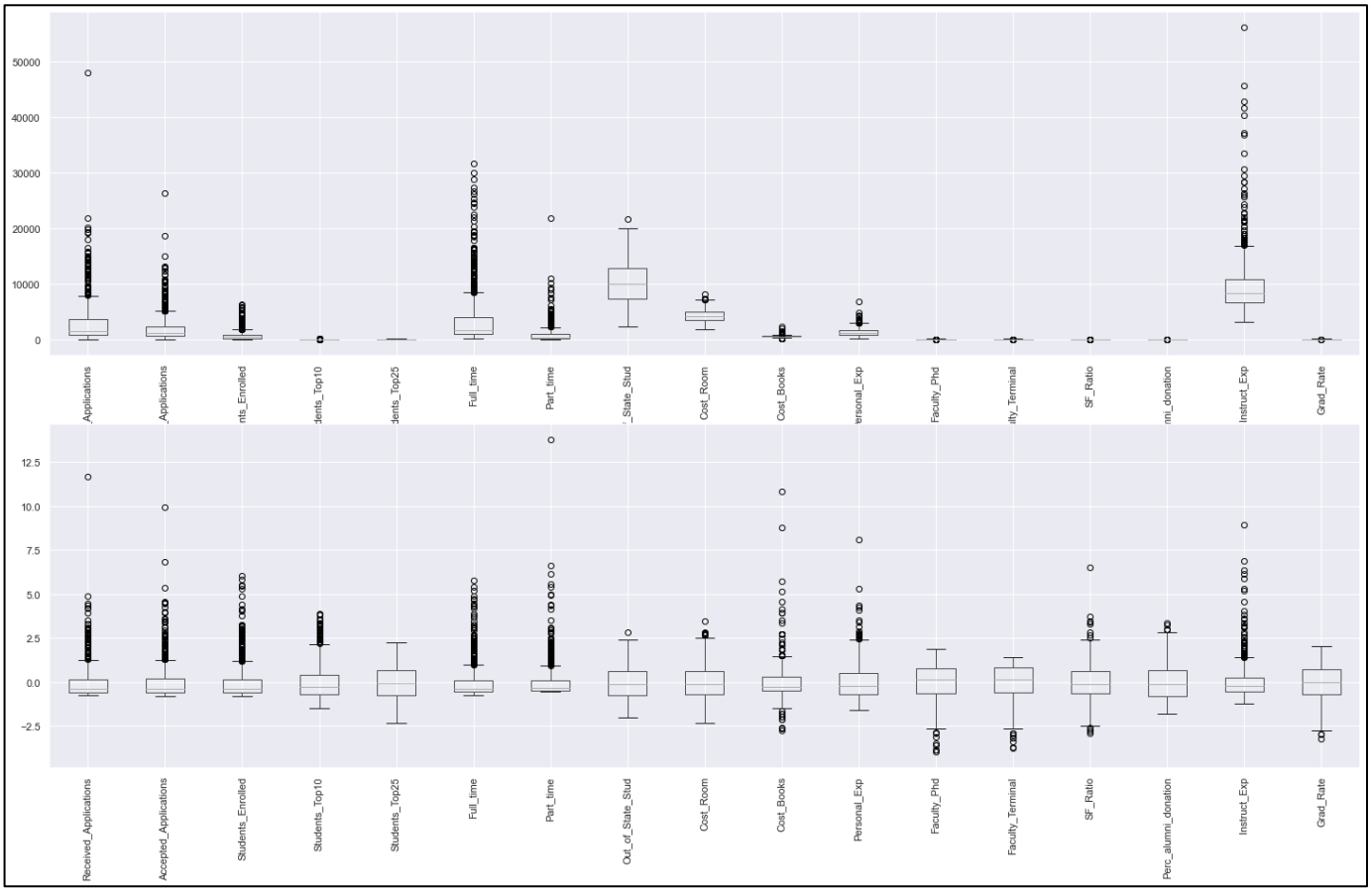
	Received_Applications	Accepted_Applications	Students_Enrolled	Students_Top10	Students_Top25	Full_time	Part_time
Received_Applications	1.00	0.94	0.85	0.34	0.35	0.81	0.40
Accepted_Applications	0.94	1.00	0.91	0.19	0.25	0.87	0.44
Students_Enrolled	0.85	0.91	1.00	0.18	0.23	0.96	0.51
Students_Top10	0.34	0.19	0.18	1.00	0.89	0.14	-0.11
Students_Top25	0.35	0.25	0.23	0.89	1.00	0.20	-0.05

The covariance matrix becomes correlation matrix for the standardized variables (after the scaling)

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

- Almost all variables have outliers.
- Students from Top 25 does not have any outliers.
- Faculty with PhD and Faculty with Terminal degree variables are left skewed.
- SF Ratio, Cost of books have outliers on both the sides (Right and left)
- Rest all columns are right skewed.
- As no outlier treatment is done, there are no changes in the presence of outliers.
- The Boxplot is scaled from -2.5 to 12.5 with 2.5 as class intervals after standardization.

<u>Graphical representation of Boxplot before and after Scaling:</u>



Statistical tests to be done before PCA:

Bartlett's test of sphericity:

It tests the hypothesis that the variables are uncorrelated in the population.

H_0: All variables in the data are uncorrelated

H_1: 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. For the given data set, p-value is less than alpha. So, we reject the null hypothesis and agree that there is correlation among variables.

Hence, PCA is recommended.

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 is less than 0.5, PCA is not recommended, since no reduction is expected.
- On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.
- For reference, Kaiser put the following values on the results:

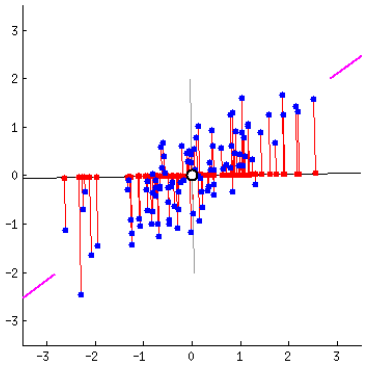
    - 0.00 to 0.49 unacceptable.
    - 0.50 to 0.59 miserable.
    - 0.60 to 0.69 mediocre.
    - 0.70 to 0.79 middling.
    - 0.80 to 0.89 meritorious.
    - 0.90 to 1.00 marvellous.

For the given data set, the measure of sampling adequacy (MSA) is 0.81, hence the data is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

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

Principal component analysis:

- Representation of data from higher dimension to lower dimension space without losing the accuracy. 2-D to 1-D.
- The objective of PCA is to detect the largest variability and convert into the first element of the coordinate system.
- The first Principal component captures most of the variability leaving the other components to pick up the residual variability.
- The Principal components are Perpendicular to each other (Orthogonal) to each other. In statistical term all the components are independent to each other.



Source: A Step-by-Step Explanation of Principal Component Analysis (builtin.com)

Steps to Perform PCA:

1. Standardization: different dimensions in the original data set, combined fairly to form the new dataset.
   X-X bar/ SD, the data is centred towards the origin. The new mean becomes '0'
2. Matrix – two dimensional representations. The covariance or correlation matrix captures relationship between different variables in the original dimensions.
3. Decompose the matrix into rotated axis. Eigen vectors - is a rotation of a data set that comes from the covariance matrix, which captures most of the variations.
4. Sort the eigen pairs in descending order ➔ identify the largest value which is considered as First Principal component ➔ this captures the maximum variance in the data points.

Data frame for 9 Principal components scores (Partial Screenshot):

	Received_Applications	Accepted_Applications	Students_Enrolled	Students_Top10	Students_Top25	Full_time	Part_time	Out_of_State_Stud	Cost_Room
0	0.25	0.21	0.18	0.35	0.34	0.15	0.03	0.29	0.25
1	0.33	0.37	0.40	-0.08	-0.04	0.42	0.32	-0.25	-0.14
2	-0.06	-0.10	-0.08	0.04	-0.02	-0.06	0.14	0.05	0.15
3	0.28	0.27	0.16	-0.05	-0.11	0.10	-0.16	0.13	0.18
4	0.01	0.06	-0.06	-0.40	-0.43	-0.04	0.30	0.22	0.56
5	-0.02	0.01	-0.04	-0.05	0.03	-0.04	-0.19	-0.03	0.16
6	-0.04	-0.01	-0.03	-0.16	-0.12	-0.03	0.06	0.11	0.21
7	-0.10	-0.06	0.06	-0.12	-0.10	0.08	0.57	0.01	-0.22
8	-0.09	-0.18	-0.13	0.34	0.40	-0.06	0.56	-0.00	0.28

2.6 Extract the eigenvalues, and eigenvectors.

Data frame for Eigen Vector:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	-0.25	0.33	-0.07	0.28	-0.00	-0.02	-0.11	0.00	-0.08	0.06	-0.17	-0.55	0.05	-0.15	0.14	-0.59	0.01
1	-0.21	0.37	-0.11	0.26	-0.05	0.01	-0.06	-0.00	-0.18	0.03	0.30	0.68	-0.07	0.19	0.06	-0.29	-0.15
2	-0.18	0.40	-0.09	0.16	0.06	-0.04	0.04	0.04	-0.13	0.03	-0.71	0.15	-0.08	-0.04	-0.12	0.45	0.03
3	-0.35	-0.08	0.04	-0.05	0.40	-0.05	-0.16	0.11	0.35	0.08	0.09	0.22	-0.02	-0.68	-0.12	-0.03	0.05
4	-0.34	-0.04	-0.01	-0.12	0.42	0.04	-0.12	0.08	0.41	0.01	-0.06	-0.11	-0.28	0.61	0.16	0.01	-0.08
5	-0.15	0.42	-0.06	0.10	0.05	-0.04	0.06	0.05	-0.07	0.01	0.60	-0.37	-0.08	-0.02	-0.12	0.50	0.07
6	-0.03	0.31	0.15	-0.15	-0.31	-0.21	0.58	0.12	0.54	-0.22	-0.03	0.05	0.10	-0.03	0.03	-0.11	-0.06
7	-0.29	-0.25	0.05	0.14	-0.22	-0.04	0.05	-0.10	-0.01	0.19	-0.01	-0.11	0.07	-0.06	-0.06	0.15	-0.83
8	-0.25	-0.14	0.14	0.20	-0.56	0.16	-0.13	-0.26	0.26	0.28	-0.00	0.03	-0.35	0.01	-0.07	0.06	0.38
9	-0.06	0.06	0.68	0.10	0.11	0.64	0.15	0.22	-0.13	-0.09	-0.01	-0.00	0.04	0.01	-0.06	-0.02	-0.03
10	0.04	0.22	0.50	-0.22	0.22	-0.32	-0.03	-0.68	-0.10	0.14	-0.00	0.01	-0.03	0.01	0.03	-0.03	-0.04
11	-0.32	0.06	-0.11	-0.54	-0.15	0.09	-0.08	-0.02	-0.17	-0.14	-0.02	-0.04	0.05	0.12	-0.67	-0.19	0.02
12	-0.32	0.05	-0.05	-0.52	-0.21	0.15	-0.04	0.03	-0.25	-0.08	0.00	0.03	-0.05	-0.17	0.66	0.12	0.02
13	0.18	0.25	-0.28	-0.17	0.06	0.49	0.01	-0.23	0.28	0.50	-0.00	0.02	0.42	0.02	0.03	0.03	-0.05
14	-0.21	-0.25	-0.15	0.01	0.22	-0.04	0.72	-0.02	-0.29	0.39	0.02	0.00	-0.14	0.02	-0.02	-0.11	0.19
15	-0.32	-0.13	0.22	0.09	-0.07	-0.30	-0.12	0.20	-0.05	0.17	0.01	0.05	0.70	0.23	0.06	0.11	0.27
16	-0.25	-0.17	-0.21	0.26	0.11	0.22	0.16	-0.53	0.05	-0.58	0.01	0.01	0.26	0.00	0.04	0.08	0.10

Rows: 17 features of the data

Columns: Principal components for the features.

Data frame for Eigen Values:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	5.45	4.49	1.17	1.0	0.94	0.85	0.59	0.6	0.52	0.41	0.02	0.03	0.31	0.09	0.14	0.16	0.22

- 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.
- Eigen vectors lead to eigen values
- 17 Eigen values represents 17 Principal components.

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only

Data frame of Percentage of variances explained by each component:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	32.06	26.4	6.9	5.91	5.52	5.0	3.55	3.44	3.07	2.38	1.83	1.28	0.95	0.84	0.53	0.2	0.14

The first feature i.e., PC-1 explains roughly 32% of the variance within the given data set. The explicit form of the PC-1 with each feature in terms of the eigen vectors are shown in the below screenshot.

	PC1
Received_Applications	0.25
Accepted_Applications	0.21
Students_Enrolled	0.18
Students_Top10	0.35
Students_Top25	0.34
Full_time	0.15
Part_time	0.03
Out_of_State_Stud	0.29
Cost_Room	0.25
Cost_Books	0.06
Personal_Exp	-0.04
Faculty_Phd	0.32
Faculty_Terminal	0.32
SF_Ratio	-0.18
Perc_alumni_donation	0.21
Instruct_Exp	0.32
Grad_Rate	0.25

Data frame showing the explicit form of PC1 corresponding to each and every data point:

	Received_Applications	Accepted_Applications	Students_Enrolled	Students_Top10	Students_Top25	Full_time	Part_time	Out_of_State_Stud	Cost_Room
0	-0.09	-0.07	-0.01	-0.09	-0.07	-0.03	-0.01	-0.22	-0.24
1	-0.05	-0.01	-0.05	-0.23	-0.47	-0.03	0.01	0.13	0.48
2	-0.10	-0.08	-0.08	-0.11	-0.10	-0.08	-0.01	0.06	-0.14
3	-0.17	-0.14	-0.12	0.65	0.58	-0.10	-0.01	0.18	0.25
4	-0.18	-0.16	-0.14	-0.23	-0.21	-0.11	0.00	-0.21	-0.05
5	-0.16	-0.13	-0.12	0.21	0.11	-0.10	-0.01	0.22	-0.23
6	-0.17	-0.14	-0.13	-0.21	-0.19	-0.10	-0.01	0.21	0.31
7	-0.07	-0.03	-0.06	0.19	0.21	-0.07	-0.01	0.25	0.11
8	-0.13	-0.10	-0.10	0.05	0.13	-0.09	-0.01	0.38	0.01

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?

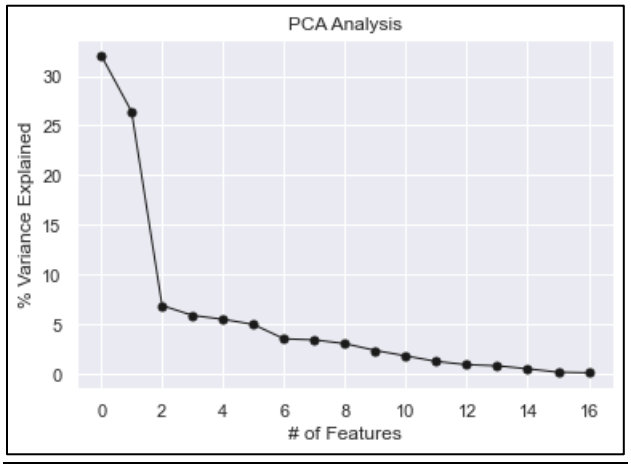
Data frame of Cumulative percentage of the explained variances by components:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	32.06	58.46	65.36	71.27	76.79	81.78	85.33	88.78	91.85	94.23	96.07	97.34	98.29	99.13	99.67	99.86	100.0

Methods for deciding on the optimum number of Principal components:

1. *Scree plot test*: A scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. The scree plot is used to determine the number of factors to retain in an exploratory factor analysis (FA) or principal components to keep in a principal component analysis (PCA). A scree plot always displays the eigenvalues in a downward curve, ordering the eigenvalues from largest to smallest. According to the scree test, the "elbow" of the graph where the eigenvalues seem to level off is found and factors or components to the left of this point should be retained as significant.
2. *Kaiser's stopping rule:* Kaiser's stopping rule Kaiser's stopping rule states that only the number of factors with eigenvalues over 1.00 should be considered in the analysis.
3. *Percent of cumulative variance:* The % of Variance column gives the ratio, expressed as a percentage, of the variance accounted for by each component to the total variance in all of the variables. The Cumulative % column gives the percentage of variance accounted for by the first n components. 90% is the cut-off threshold.

Scree plot test



According to Scree test we will take only first *3 components* for Principal component analysis.

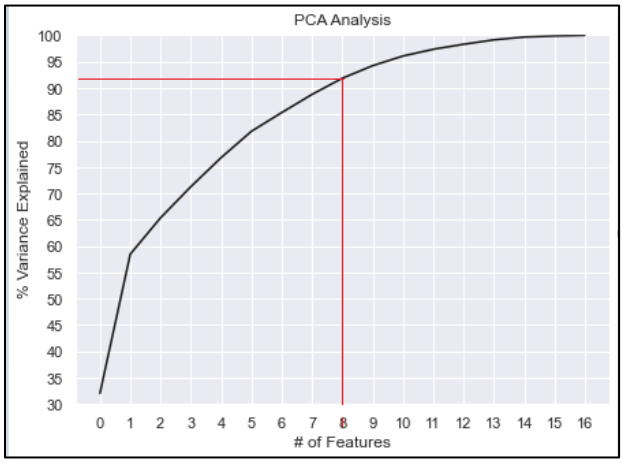
Kaiser's stopping rule:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	5.45	4.49	1.17	1.0	0.94	0.85	0.59	0.6	0.52	0.41	0.02	0.03	0.31	0.09	0.14	0.16	0.22

There are *4 components* with eigen values over 1.00

Percent of cumulative variance:

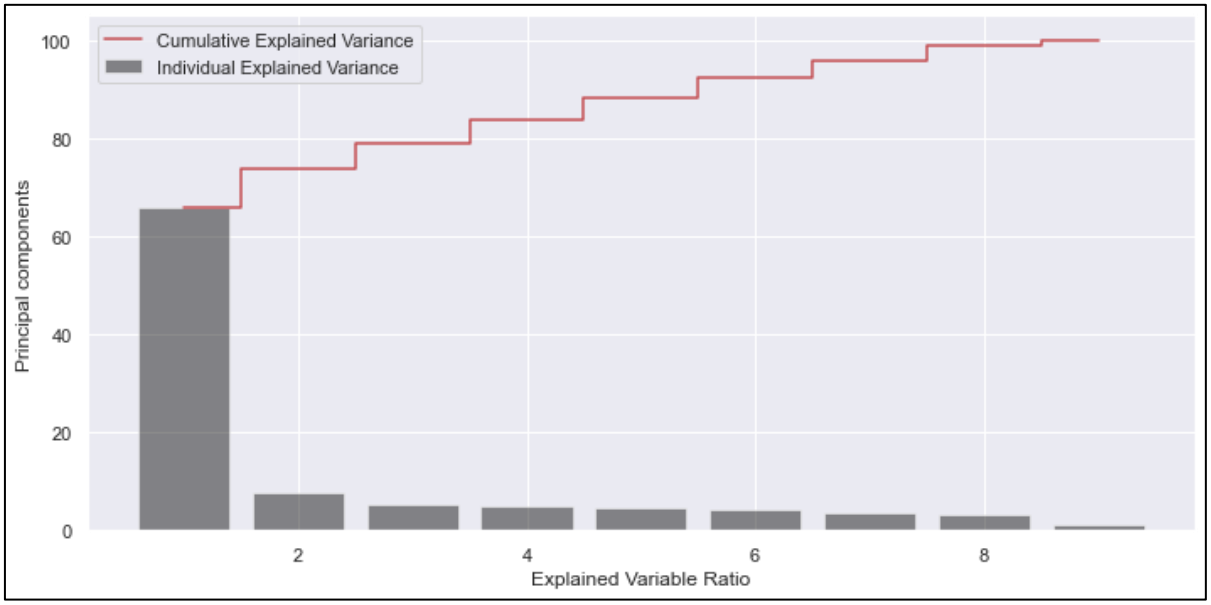
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	32.06	58.46	65.36	71.27	76.79	81.78	85.33	88.78	91.85	94.23	96.07	97.34	98.29	99.13	99.67	99.86	100.0



As 90% is the threshold for selecting the cumulative percent, we will choose 9 components explain 90% approximately.

For this particular data set, to select an institution for graduation, Graduation rate, SF Ratio, Faculty quality, student quality, expenses involved are the key elements. Considering only 3 to 4 components will not suffice as they explain only 70% of variation. When more variations are compressed, the PCA works well. So, the optimum number of components is 10.

Graphical representation: Plot showing the Explained variance and Cumulative percentage variance:



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]

With the help of the first Principal component, we are analysing the 17 features of the data to find the top 10 colleges to graduate. A new column is added to the data frame called 'Univ/College_Score'. This column shows the total score of the colleges to decide which College is best to graduate from. Higher the score the College is marked the best.

The below data frame shows the 15 best Universities to graduate. This interpretation is based on PC1 component.

	Univ/College	Univ/College_Score
284	Johns Hopkins University	8.047182
775	Yale University	7.919327
250	Harvard University	7.694199
483	Rutgers at New Brunswick	7.320976
663	University of Pennsylvania	7.177764
174	Duke University	7.140504
728	Washington University	6.932285
459	Princeton University	6.927871
158	Dartmouth College	6.853530
354	Massachusetts Institute of Technology	6.741215
637	University of Michigan at Ann Arbor	6.687848
70	Brown University	6.516179
424	Northwestern University	6.509921
609	University of Chicago	6.409190
191	Emory University	6.267502

2021 Best Colleges Rankings by US-News have listed University of Michigan at Ann Arbor, Massachusetts Institute of Technology (MIT), Harvard University, University of Chicago , University of California at Irvine, University of California at Berkeley and Columbia University as the best universities to graduate. The best colleges arrived after implementing PCA also considered ranked as Best Universities by US-News.

Interpretations from the Summary of the Top 3 Institutions:

1. The average number of received applications is more than 11,000 and the maximum number of applications are 14000 approximately.
2. The student's enrolment is also high.
3. The count of students from Top10 and Top 25 are significantly high. Hence, *the statement of students from Top 10 and Top 25 have high graduation rate is clearly proved.*
4. Out of state students count is also significantly high for the top 3 colleges.
5. Almost 100 faculties of these institutions hold either PhD or Terminal degree. Hence, the finding of *intuitions with high number of PhD and Terminal degree faculties have high graduation rate is proved.*
6. When SF Ratio is low the Graduation rate is high. Hence, institutions with low SF Ratio are considered the best.
7. The instructional expenses are high for the best institutions.
8. Percentage of Alumni donation is high.
9. Cost of books are very affordable and it very close to the mean cost of books.

10. The average graduation rate is 96% and the maximum is 100%. Universities with significantly high graduation rate are ranked the best.
11. The highest score computed by analysing only the first Principal component is 8.05 which pertains to ***John Hopkins University***

The summary of the top 3 Universities is explained using the description function. Below is the screenshot of the summary:

	count	mean	std	min	25%	50%	75%	max
Received_Applications	3.0	11014.67	2708.81	8474.00	9589.50	10705.00	12285.00	13865.00
Accepted_Applications	3.0	2688.00	672.06	2165.00	2309.00	2453.00	2949.50	3446.00
Students_Enrolled	3.0	1278.00	349.14	911.00	1114.00	1317.00	1461.50	1606.00
Students_Top10	3.0	86.67	10.41	75.00	82.50	90.00	92.50	95.00
Students_Top25	3.0	97.67	3.21	94.00	96.50	99.00	99.50	100.00
Full_time	3.0	5215.00	1648.00	3566.00	4391.50	5217.00	6039.50	6862.00
Part_time	3.0	657.33	798.37	83.00	201.50	320.00	944.50	1569.00
Out_of_State_Stud	3.0	19041.67	709.09	18485.00	18642.50	18800.00	19320.00	19840.00
Cost_Room	3.0	6553.33	169.21	6410.00	6460.00	6510.00	6625.00	6740.00
Cost_Books	3.0	543.33	75.06	500.00	500.00	500.00	565.00	630.00
Personal_Exp	3.0	1691.67	572.72	1040.00	1480.00	1920.00	2017.50	2115.00
Faculty_Phd	3.0	96.33	0.58	96.00	96.00	96.00	96.50	97.00
Faculty_Terminal	3.0	96.67	0.58	96.00	96.50	97.00	97.00	97.00
SF_Ratio	3.0	6.33	3.33	3.30	4.55	5.80	7.85	9.90
Perc_alumni_donation	3.0	46.33	7.37	38.00	43.50	49.00	50.50	52.00
Instruct_Exp	3.0	44612.67	10187.32	37219.00	38802.50	40386.00	48309.50	56233.00
Grad_Rate	3.0	96.33	5.51	90.00	94.50	99.00	99.50	100.00
Univ/College_Score	3.0	7.89	0.18	7.69	7.81	7.92	7.98	8.05

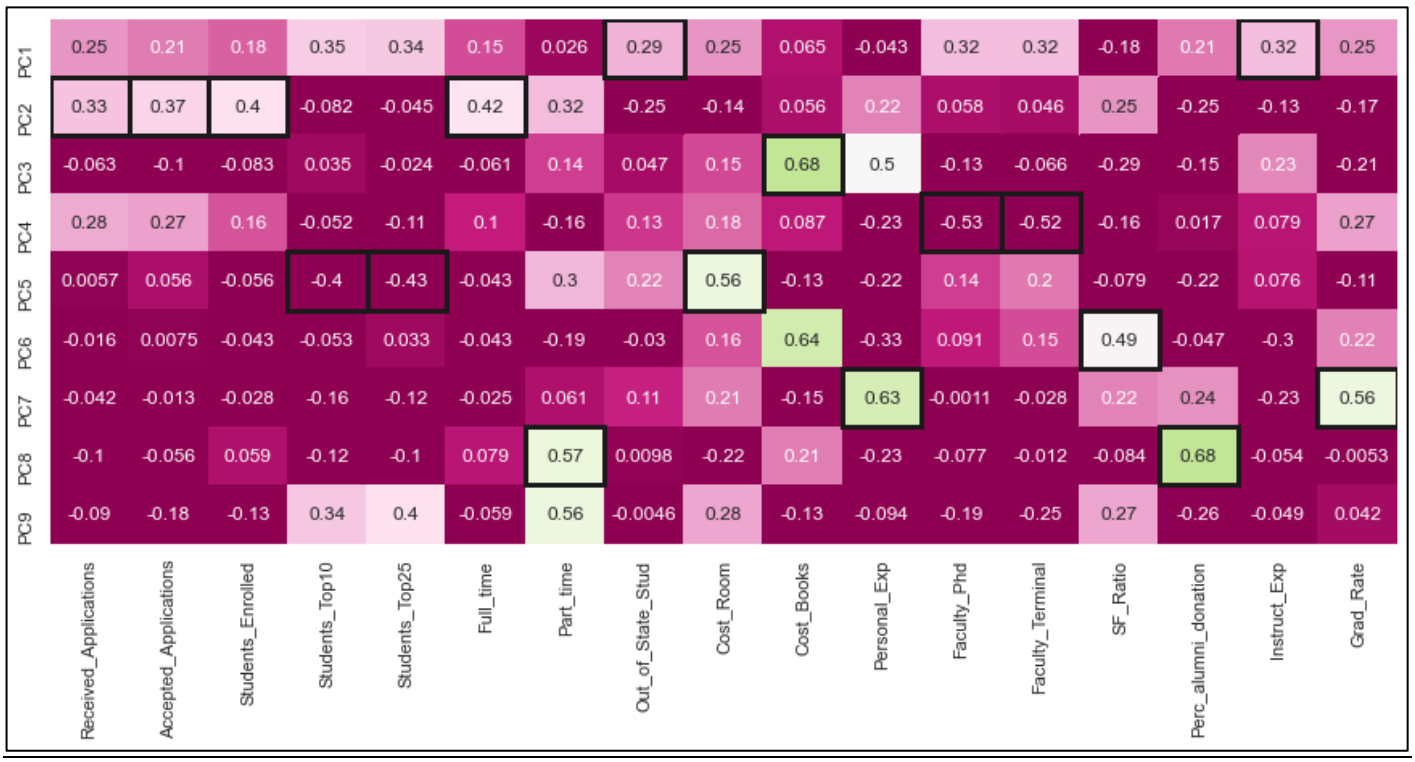
The summary of the last 3 Universities with least score as per PCA is explained using the description function. Below is the interpretation of the summary and its screenshot.

Interpretations from the Summary of the last 3 Institutions:

1. Less number of received applications when compared with institutions with high score.
2. Students from Top 10 and 25 Higher secondary class are less.
3. Average of 40-50 faculties hold PhD and Terminal degrees.
4. The Student Faculty ratio is high. 23 students for each faculty.
5. Percentage of alumni donation is low.
6. The graduation rate is below 50%
7. The least score is -3.84
8. Cost of Books are comparatively higher than the mean cost of books which is 500.

	count	mean	std	min	25%	50%	75%	max
Received_Applications	3.0	572.33	213.18	331.00	491.00	651.00	693.00	735.00
Accepted_Applications	3.0	445.00	126.44	331.00	377.00	423.00	502.00	581.00
Students_Enrolled	3.0	278.00	76.74	225.00	234.00	243.00	304.50	366.00
Students_Top10	3.0	14.33	6.03	8.00	11.50	15.00	17.50	20.00
Students_Top25	3.0	33.67	15.63	17.00	26.50	36.00	42.00	48.00
Full_time	3.0	1388.33	948.94	617.00	858.50	1100.00	1774.00	2448.00
Part_time	3.0	302.33	356.61	34.00	100.00	166.00	436.50	707.00
Out_of_State_Stud	3.0	7016.67	2110.55	5000.00	5920.00	6840.00	8025.00	9210.00
Cost_Room	3.0	3717.33	66.04	3650.00	3685.00	3720.00	3751.00	3782.00
Cost_Books	3.0	800.00	264.58	600.00	650.00	700.00	900.00	1100.00
Personal_Exp	3.0	2171.00	2383.05	600.00	800.00	1000.00	2956.50	4913.00
Faculty_Phd	3.0	42.33	8.33	33.00	39.00	45.00	47.00	49.00
Faculty_Terminal	3.0	43.00	9.17	33.00	39.00	45.00	48.00	51.00
SF_Ratio	3.0	23.07	14.51	14.00	14.70	15.40	27.60	39.80
Perc_alumni_donation	3.0	14.33	6.03	8.00	11.50	15.00	17.50	20.00
Instruct_Exp	3.0	6837.33	1470.46	5524.00	6043.00	6562.00	7494.00	8426.00
Grad_Rate	3.0	42.67	7.77	34.00	39.50	45.00	47.00	49.00
Univ/College_Score	3.0	-3.90	0.06	-3.95	-3.93	-3.91	-3.87	-3.84

Graphical representation of other Principal components explaining variances in key features:



Data frame formed using reduced data from the 9 Principal components (Partial Screenshot):

The 9 principal components explain variances of different features. Below is the screenshot of the data frame and its summary.

Univ/College	PC1_Instruct_Exp	PC2_FullTime	PC3_Cost_Books	PC4_Faculty_Phd_Terminal	PC5_Cost_Room	PC6_SFRatio	PC7_Grad_Rate
Abilene Christian University	-1.59	0.77	-0.10	-0.92	-0.74	-0.30	0.64
Adelphi University	-2.19	-0.58	2.28	3.59	1.06	-0.18	0.24
Adrian College	-1.43	-1.09	-0.44	0.68	-0.37	-0.96	-0.25
Agnes Scott College	2.86	-2.63	0.14	-1.30	-0.18	-1.06	-1.25

	count	mean	std	min	25%	50%	75%	max
PC1_Instruct_Exp	777.0	0.0	2.33	-5.66	-1.73	-0.30	1.34	8.05
PC2_FullTime	777.0	0.0	2.12	-3.59	-1.35	-0.63	0.69	12.00
PC3_Cost_Books	777.0	0.0	1.08	-2.94	-0.67	-0.10	0.49	9.01
PC4_Faculty_Phd_Terminal	777.0	-0.0	1.00	-2.94	-0.66	-0.06	0.60	5.18
PC5_Cost_Room	777.0	0.0	0.97	-2.69	-0.70	-0.05	0.63	4.25
PC6_SFRatio	777.0	-0.0	0.92	-3.82	-0.52	-0.00	0.46	5.99
PC7_Grad_Rate	777.0	0.0	0.78	-2.81	-0.51	0.04	0.47	4.35
PC8_Alumni_Donation	777.0	-0.0	0.77	-1.85	-0.47	-0.04	0.42	8.43
PC9_Top10_25	777.0	0.0	0.73	-2.60	-0.46	-0.05	0.39	5.25

Conclusion:

With help of PCA we have been able to reduce 17 numeric features into 9 components which is able to explain 90% of variance in the data

With help of reduced components, we have been able to observe some patterns. Using the first principal component we analysed the Universities with high scores and low scores. The Universities with the high scores are cross checked with the latest Rankings by US-News. We are able to find the best universities of our list in the Ranking list as well.

Unsupervised learning like clustering can further be applied on the data to segment the customers based on the components created and further analysed.

Libraries imported in the Jupyter codebook:

Problem:1

```
import os
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy .stats as stats
%matplotlib inline
sns.set()
import math
from scipy import stats
from statsmodels.formula.api import ols
from statsmodels.stats.anova import _get_covariance,anova_lm
from statsmodels.graphics.factorplots import interaction_plot
from statsmodels.stats.multicomp import (pairwise_tukeyhsd,
                                         MultiComparison)
import statsmodels.stats.multicomp as mc
```

Problem:2

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import math
from scipy.stats imp import ort zscore
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
from factor_analyzer.factor_analyzer import calculate_kmo
from matplotlib.patches import Rectangle
```