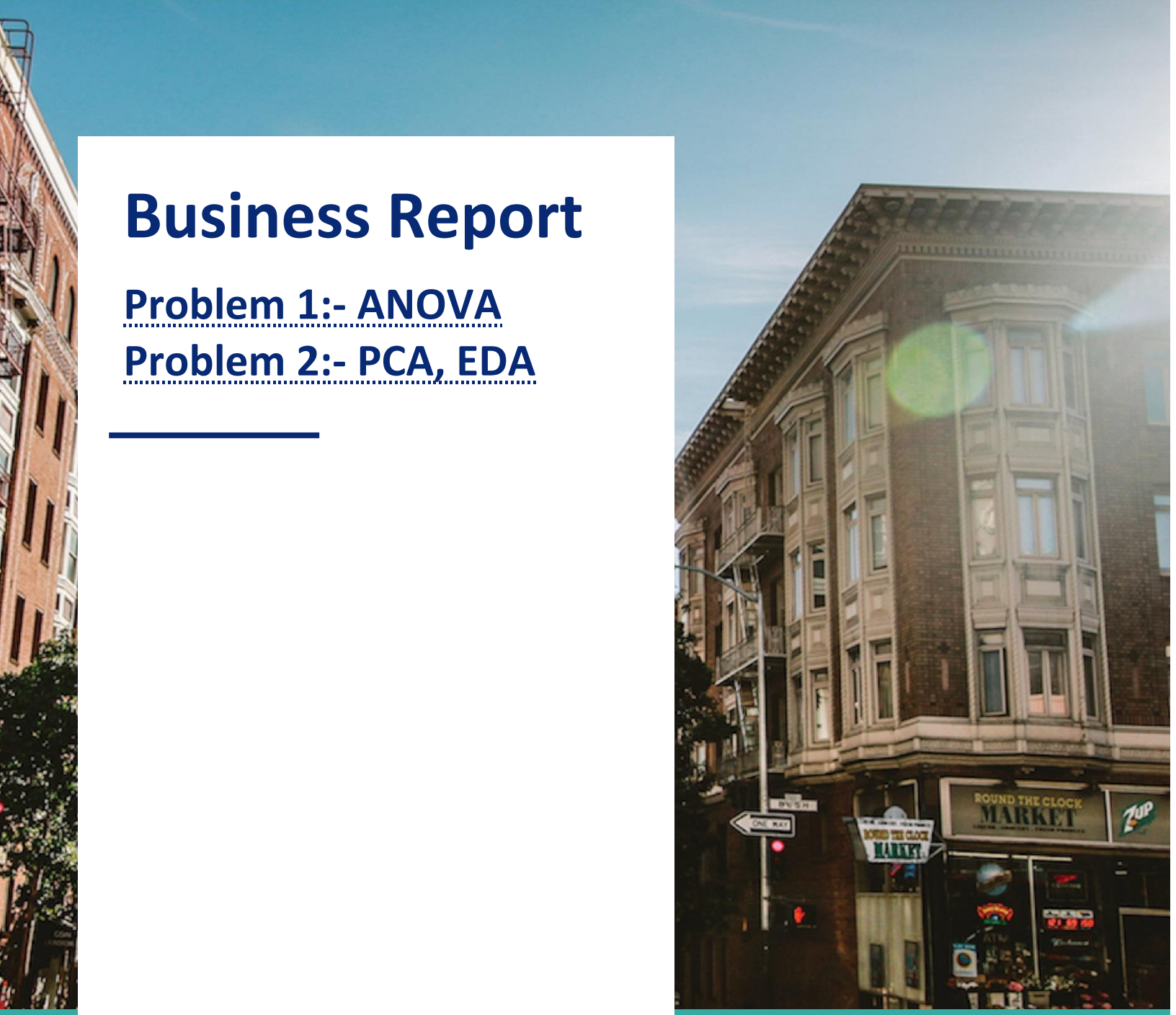


Business Report

Problem 1:- ANOVA

Problem 2:- PCA, EDA

MAY 12 2021

HARSH ALKESH PANDYA

PGP-DSBA Online Feb_A' 2021

INDEX

Table of Content

Summary
Introduction
Problem 1 1-6
1.1..... 1
1.2..... 2
1.3..... 3
1.4..... 3
1.5..... 4
1.6..... 5
1.7..... 6

Problem 2.....7-28
2.1.....7
2.2.....12
2.3.....13
2.4.....16
2.5.....18
2.6.....20
2.7.....22
2.8.....23
2.9.....25-28

List of Figures

1.....4
2.....7
3.....8
4.....9
5.....10
6.....17
7.....17
8.....23
9.....24
10.....25

List of Tables

1.....2
2.....2
3.....3
4.....5
5.....12
6.....15
7.....15

SUMMARY

Problem 1: -

Salary

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, the normality assumption may not always hold if the sample size is small.]

Problem 2: -

Education

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.

TERMS

Prob 1A

1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.
2. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
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.
4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. **(Non-Graded)**

Prob 1B

1. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the 'pointplot' function from the 'seaborn' function]
2. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?
3. Explain the business implications of performing ANOVA for this particular case study.

Prob 2

- Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
- Is scaling necessary for PCA in this case? Give justification and perform scaling.
- Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
- Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]
- Extract the eigenvalues and eigenvectors.[print both]
- Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features
- Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).
- 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?
- 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]

Problem 1

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

Solution: -

As mention on notes We Assume that the data follows a **Normal Distribution**

So, this is what we call alpha (α). The typical value of **α = 0.05**. This means that there is 95% confidence that the conclusion of this test will be valid.

Normal Distribution - A variable is said to be normally distributed or have a normal distribution if its distribution has the shape of a normal curve a special bell-shaped curve**.** The graph of a normal distribution is called the normal curve**,** which has all the following properties: - The mean, median, and mode are equal.

The hypothesis of the One-way ANOVA of "Salary" variable with the ingredient "Education" variable.

- H0: The means of 'Salary' variable with respect to each amount of ingredient 'Education' is equal
- H1: At least one of the means of the 'Salary' variable with respect to each amount of ingredient 'Education' is unequal.

The hypothesis of the One-way ANOVA of "Salary" variable with the ingredient "Occupation" variable.

- H0: The means of 'Salary' variable with respect to each amount of ingredient 'Occupation' is equal
- H1: At least one of the means of the 'Salary' variable with respect to each amount of ingredient 'Occupation' is unequal.

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.

Solution: -

Output: -

	Source	SS	DF	MS	F	p-unc	np2
0	Education	1.026955e+11	2	5.134773e+10	30.95628	1.257709e-08	0.625932
1	Within	6.137256e+10	37	1.658718e+09	NaN	NaN	NaN

```
1.257709e-08<0.05
```

```
True
```

(Table 1)

Conclusion: - p-value is smaller than chosen alpha level α = 0.05, so null hypothesis can be rejected.

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.

Solution: -

Output: -

	Source	SS	DF	MS	F	p-unc	np2
0	Occupation	1.125878e+10	3	3.752928e+09	0.884144	0.458508	0.068623
1	Within	1.528092e+11	36	4.244701e+09	NaN	NaN	NaN

```
0.458508<0.05
```

```
False
```

(Table 2)

Conclusion: - p-value is greater than chosen alpha level α = 0.05, so the null hypothesis cannot be rejected. Hence, we accept the Null Hypothesis.

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.

Solution: -

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	2.284576	9.648715e-02
C(Education)	2.0	9.695663e+10	4.847831e+10	29.510933	3.708479e-08
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

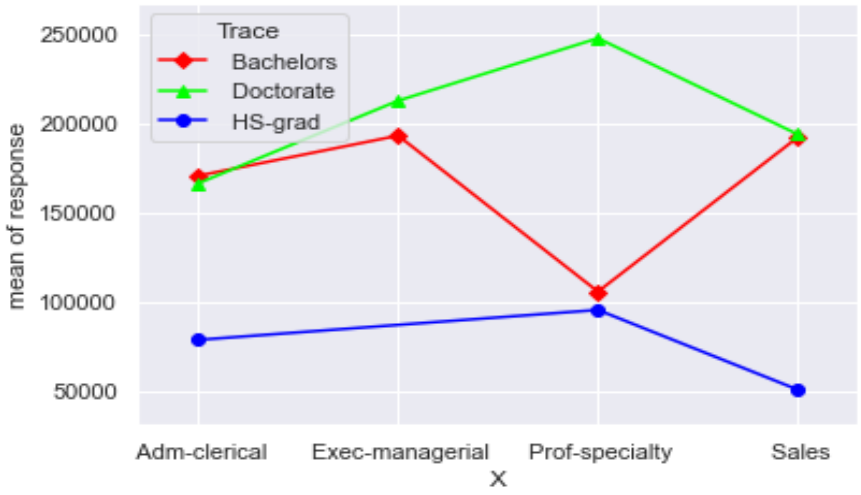
(Table 3)

As shown in table Education null hypophysis can be rejected. by P-value is higher than the significance level The Education and Occupation Both of the Mean are significantly Different with each other. The integration result said that the class of means are different and both the p value is less then the level of significance so both terms are logically reject the null hypothesis.

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.

Solution: -

Analysis of the effects of one variable on another with the help of an interaction plot.



(Fig 1: Interaction plot)

Insights from the graphs (Interaction plot):

- The second amount variation for both ingredients Education and Occupation have very strong interaction with the third variation.
- In the case of Ingredient, Occupation and Volunteer Education, the highest interaction for Bachelors and Doctorate amount variation can be observed while amount HS-grade has no interaction from other amount variation. The relief is also the lowest from amount HS-grade
- We can observe the strong evidence of interaction between the variation Bachelors and Doctorate. For volunteer Doctorate the interaction is strongest and the difference in terms of relief is less than one. While looking into amount HS-grade of ingredient Education the interaction is weakest.

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?

Solution: -

H0: The means 'Salary' time with respect to each level of compound Education and Occupation is equal. (There is an interaction effect)

HA: At least one of the means of 'Salary' time with respect to each level of compound Education and Occupation is unequal.

If the interaction p-value is statistically significant, then we conclude that the effect on the mean outcome of a change in one factor depends on the level of the other factor. More specifically, for at least one pair of levels of one factor the effect of a particular change in levels for the other factor depends on which level of the first pair we are focusing on.

Conclusion: - If p-value < significance level (0.05); Reject the null hypothesis (plays significant role because means are different)

Output: -

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	5.277862	4.993238e-03
C(Education)	2.0	9.695663e+10	4.847831e+10	68.176603	1.090908e-11
C(Occupation):C(Education)	6.0	3.523330e+10	5.872217e+09	8.258287	2.913740e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

(Table 4)

- Here the p value for Occupation > 0.05, which means we to fail reject the null hypothesis for it. It does not play a significant role in determining the mean Salary for the Occupation.
- The p value for Education < 0.05, which means we reject the null hypothesis for it. It plays a significant role in determining the mean Salary for the Education.
- The p value for interaction of Education and Occupation < 0.05, which means we reject the null hypothesis for it. There is no conclusive evidence of there being an interaction between compounds Education and Occupation.

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

Solution: -

The analysis of the variables of the given data and the respective results leads to two significant conclusions on a business level. These conclusions are as follows:

From the previous questions, we have deduced that there is not significant interaction between compounds Occupation and Education, and that the relief times of both compounds are same. We have also come to know from two-way ANOVA that compound Occupation plays a significant role in determining the mean relief time for the relief of severe cases of hay fever than compound Education. Hence it should be preferred more.

Now coming to which level provides better relief time, we run Two-way test for Occupation and Education (both come to be Different). We can derive by looking at the interaction graph that the Salary of Doctorate at high level provides the most Salary of Occupation from Education is Prof-Specialty as compared to Education and Occupation statistically, by looking at the table it is said that the difference between the mean relief times of both treatments are not significant. So, both the levels of concentrations can be selected.

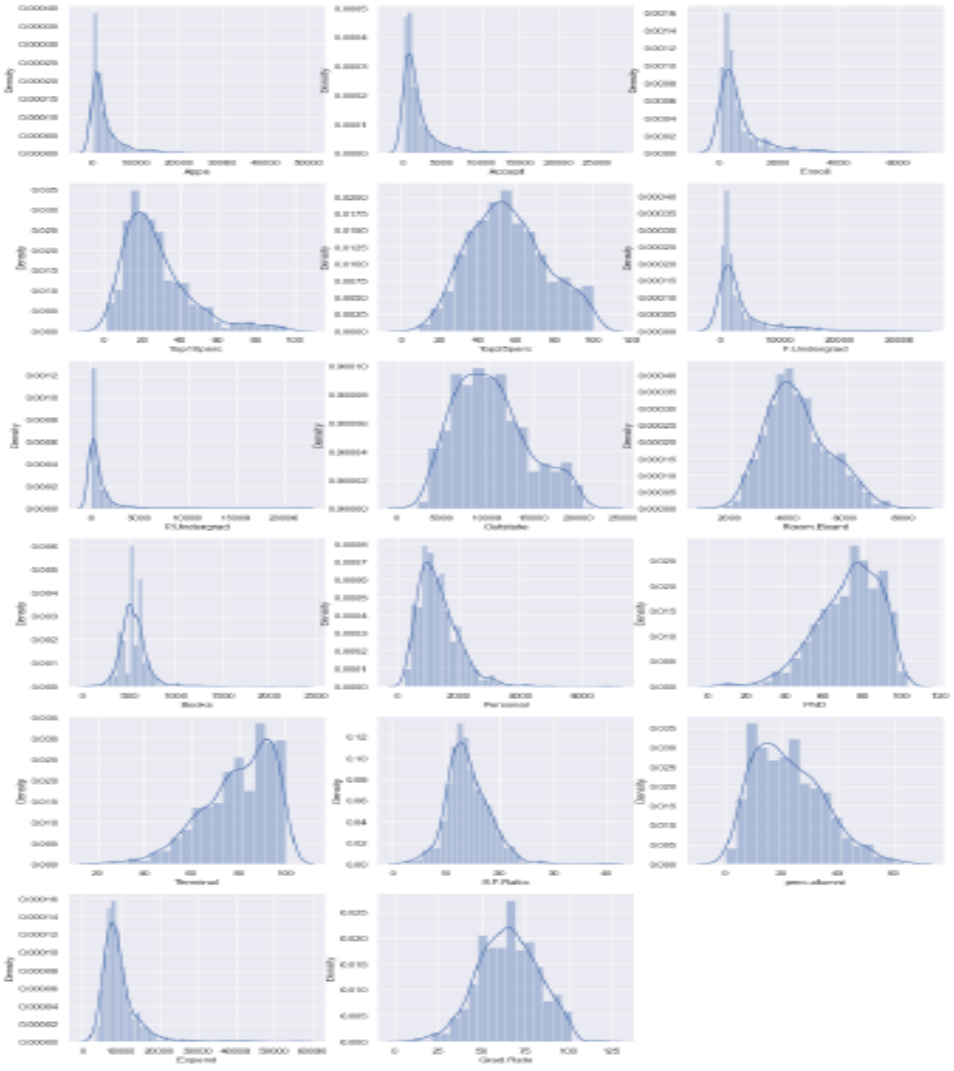
Problem 2

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

Solution: -

Univariate analysis

We will check the distribution of each variable by plotting their histogram:

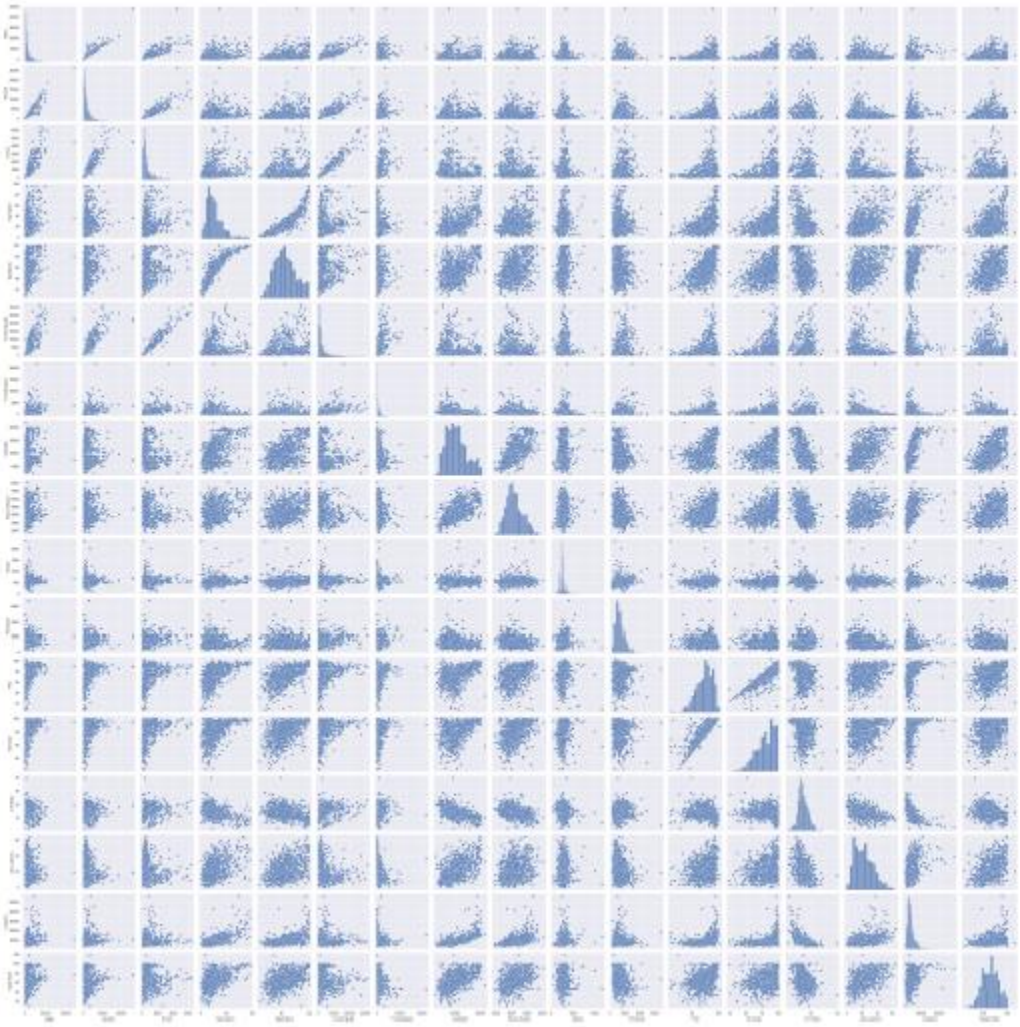


(Fig 2 :- Univariate histography)

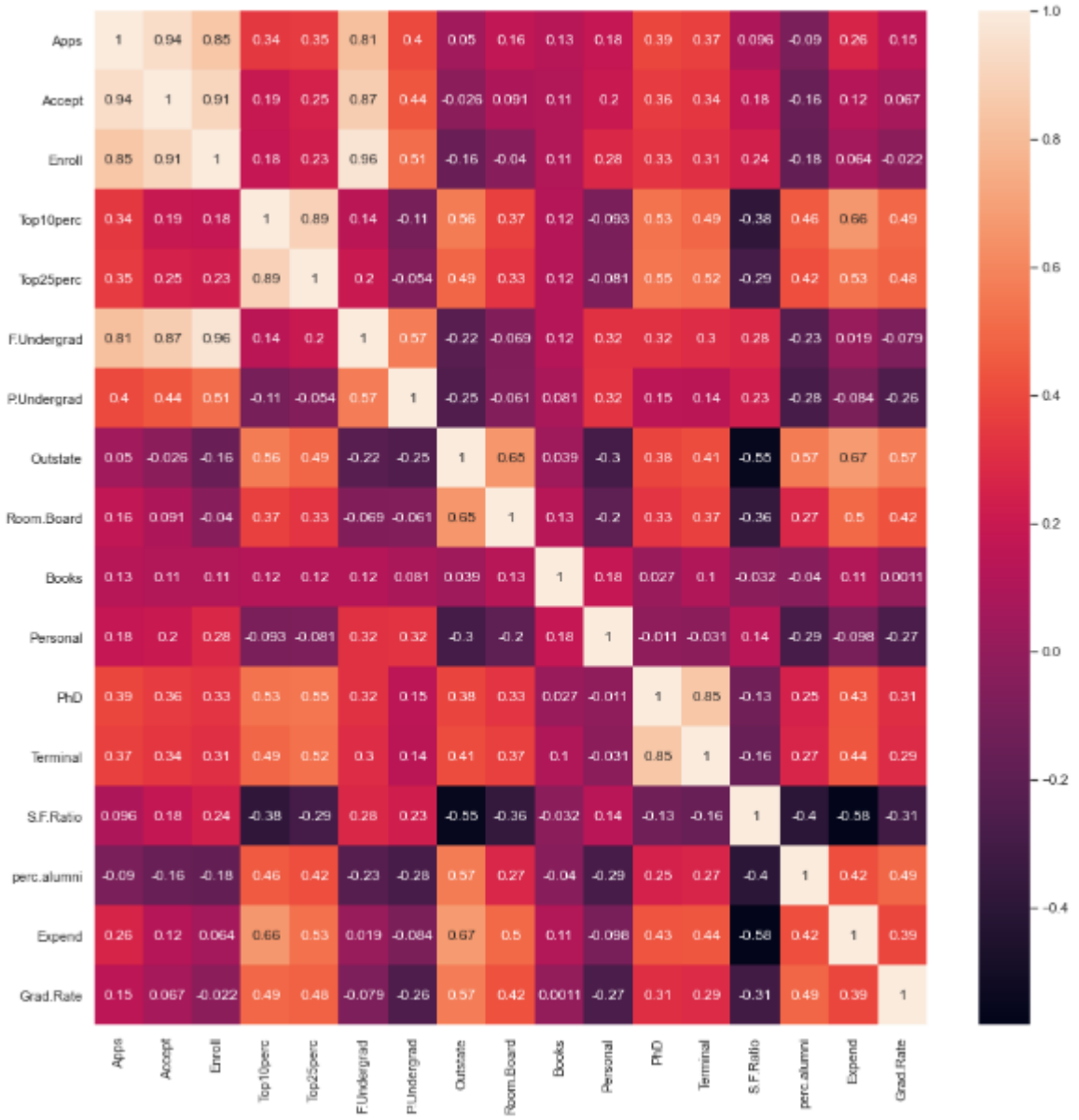
We observe

- By looking at the plot we observe that there are clearly describe; It takes data, summarizes that data and finds patterns in the data.
- Variables like Top25perc Distribution, Outstate Distribution, Room.Board Distribution, perc.alumni Distribution and Grad.Rate Distribution appear to be almost normally distributed which also result in fewer outliers than others.
- ∉ Most of the variables are right skewed, i.e., they show positive skewness, and their tail ends in right.

Multivariate analysis



(Fig 3 :- Multivariate pair plot)



(Fig 4 :- Heatmap)

Due to Multivariate has so many data and it looks difficult to identify we take help of heat map

We start the Multivariate analysis by plotting a heat map which shows the correlation between variables. The darkest and the lightest color signify high amount of correlation.

There is a high positive correlation between apps (Number of applications received), accept (Number of applications accepted), enroll (Number of new students enrolled), F. Undergrad (Number of full-time undergraduate students). We can infer that in most colleges the students enrolled majorly for full time

graduation program.

There is a high positive correlation between PhD (Percentage of faculties with Ph.D.'s) and Terminal (Percentage of faculties with terminal degree). We can infer that the faculties having PhD is their terminal degree indeed.

There is a high positive correlation between Top10perc (Percentage of new students from top 10% of Higher Secondary class) and Top25perc (Percentage of new students from top 25% of Higher Secondary class). We can infer that the new students from top 25% also belong to the top 10% of Higher Secondary class.

(Fig 5 :- Subplot)

For analysis purpose we try to put a regression line through following variables and observe that there exists a positive linear relationship among them. As the number of apps increases, the number of accept increases, and as the number of accept increases, the number of enroll also increases.

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

Solution: -

Yes, for this case scaling is necessary for PCA in the given dataset, although all the values are numeric, they differ in scales. Some provide count of students like number of applications received, accepted and enrolled, some denote the cost of amenities like room. Board, books, expend which account for the higher numbers in the given dataset and some are just percentage values like S.F. Ratio and grad ratio whose values will always be below hundred.

PCA effectiveness depends upon the scales of the attributes. If data is having different scales, PCA has following drawbacks:

- PCA will pick variable with highest variance rather than picking up attributes based on correlation
- Changing scales of the variables can change the PCA
- Interpreting PCA can become challenging due to presence of discrete data
- Presence of skew in data with long thick tail can impact the effectiveness of the PCA (related to point 1)

Because it's trying to capture the total variance in the set of variables, PCA requires that the input variables have similar scales of measurement. Variables whose numbers are just larger will have much bigger variance just because the numbers are so big. So before starting with a covariance matrix, it's a good idea to standardize those variables before we begin so that the variables with the biggest scales don't overwhelm the PCA.

Standardization transforms the data such that the resulting distribution has a **mean of 0** and a **standard deviation of 1**. We should standardize the variables before applying PCA because it will give more emphasis to those variables having higher variances than to those variables with very low variances while identifying the right principle component.

Normalizing the data is sensitive to outliers, so if there are outliers in the data set it is a bad practice. Standardization creates a new data not bounded (unlike normalization).

Usually, the standardization (Z-score normalization) is preferred because normalization (min-max scaling) is prone to over-fitting.

- **Standardization** is mostly used on **unsupervised learning algorithms**. In our case, standardization is **more beneficial** than normalization.
- If you see a **bell-curve** in your data then **standardization** is more preferable. In our case we observed while doing uni-variate analysis that our data does contain bell curve formation hence **standardization** is preferred.
- If your dataset has **extremely high** or **low values** (**outliers**) then **standardization** is more preferred because usually, normalization will **compress** these values into a **small range**.It's also not influenced by maximum and minimum values in our data so if our data contains outliers it's good to go.

We thus standardize the data. Data on all the dimensions are subtracted from their means to shift the data points to the origin, i.e. the data is centered on the origins so that when the axis is rotated, the values don't change.

Here we use zscore method from scipy library that computes the relative z-score of the input data, relative to the sample mean and standard deviation. It computes zscores for all columns at once.

The dataset after using StandardScaler or z-score scaling :

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R.
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	1.013
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	-0.477
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	-0.300
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	-1.615
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	-0.553

(Table 5)

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

Solution: -

Both covariance and correlation measure linear relationships between variables.

The covariance indicates the direction of the linear relationship between variables. A covariance of 0 indicates that two variables are totally unrelated. If the covariance is positive, the variables increase in the same direction, and if the covariance is negative, the variables change in opposite directions. The magnitude of the covariance depends on the scale of each variable. Correlation on the other hand measures both the strength and direction of the linear relationship between two variables. When the correlation coefficient is positive, an increase in one variable also results in an increase in the other. When the correlation coefficient is negative, the changes in the two variables are in opposite directions. When there is no relationship, there is no change in either.

The linear correlation between two features and is closely related to the covariance. In fact, correlation between two variables is a normalized version of the covariance. By dividing the covariance by the features' standard deviations, we ensure that the correlation between two features is in the range [-1, 1], which makes it more interpretable than the unbounded covariance.

In fact, it's a normalized version of the covariance However, note that the covariance and correlation are the same if the features are normalized to unit variance (e.g., via standardization or z- score normalization). **Two features are perfectly positively correlated if $\rho=1\rho=1$ and perfectly negatively correlated if $\rho=-1\rho=-1$. No correlation is observed if $\rho=0\rho=0$.**

Thus, we can state that above three approaches yield the same eigenvectors and eigenvalue pairs:

- Eigen decomposition of the covariance matrix after standardizing the data.
- Eigen decomposition of the correlation matrix.
- Eigen decomposition of the correlation matrix after standardizing the data.

Finally, we can say that after scaling - the covariance and the correlation matrix have the same values.

Below is comparison between covariance matrix and correlation matrix on scaled data. We can see that they have same values.

Covariance Matrix

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
   0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
   0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
   0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
   0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
  -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
   0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.19970167
  -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
   0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.00128866
   0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
   0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
   1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
   0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
  -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
   0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
  -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
   0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
 [ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
   0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
   0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
 [ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
   0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
  -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
 [ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
   0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
   0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
 [ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
   0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
   1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
 [ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
   0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
  -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
 [-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
  -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
   0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
 [ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
  -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
   0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
 [ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
  -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
   0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

(Matrix 1)

Correlation Matrix

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Term
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	0.369
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	0.337
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	0.308
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	0.491
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	0.524
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	0.300
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	0.141
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	0.407
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	0.374
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	0.099
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	-0.030
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	0.849
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	1.000
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	-0.160
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	0.267
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	0.438
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	0.289

(Table 6)

Compare Correlation Matrix and Covariance matrix

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Term
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	0.369
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	0.337
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	0.308
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	0.491
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	0.524
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	0.300
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	0.141
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	0.407
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	0.374
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	0.099
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	-0.030
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	0.849
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	1.000
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	-0.160
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	0.267
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	0.438
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	0.289

(Table 7)

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

Solution: -

From the Bellow box plots, we can observe that the original numeric information is on different scales, like outstate and expend had values way higher than other variables.

PCA effectiveness depends upon the scales of the attributes. If data is having different scales, PCA has following drawbacks:

- PCA will pick variable with highest variance rather than picking up attributes based on correlation
- Changing scales of the variables can change the PCA
- Interpreting PCA can become challenging due to presence of discrete data
- Presence of skew in data with long thick tail can impact the effectiveness of the PCA (related to point 1)

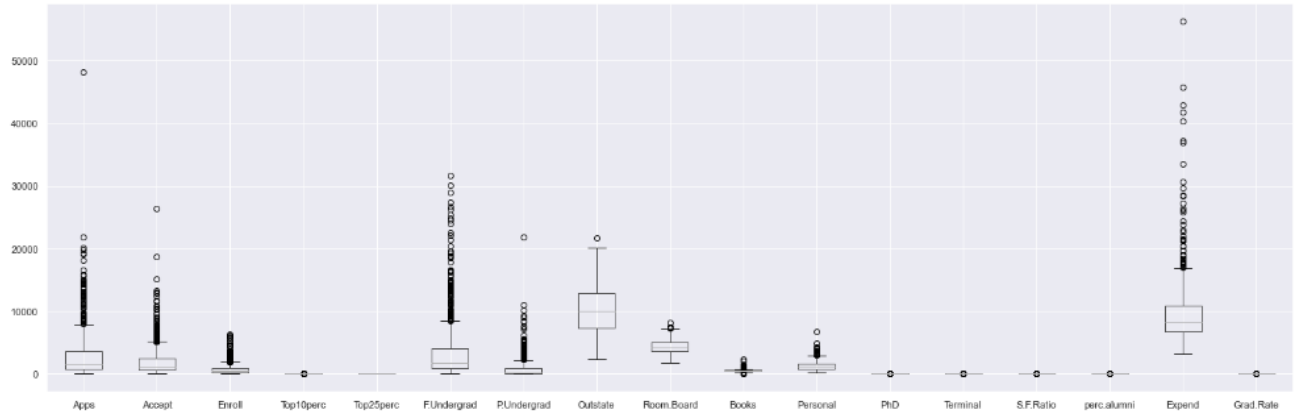
Because it's trying to capture the total variance in the set of variables, PCA requires that the input variables have similar scales of measurement. Variables whose numbers are just larger will have much bigger variance just because the numbers are so big. So before starting with a covariance matrix, it's a good idea to standardize those variables before we begin so that the variables with the biggest scales don't overwhelm the PCA. So, after scaling the variables, we can observe in box plot 2 that the variables have been standardized and are now on same scale, and now this data can be worked upon for PCA.

Don't drop an outlier if:

- **Your results are critical**, so even small changes will matter a lot. For example, you can feel better about dropping outliers about people's favorite TV shows, not about the temperatures at which airplane seals fail.
- **There are a lot of outliers.** Outliers are rare by definition. If, for example, 30% of your data is outliers, then it actually means that there's something interesting going on with your data that you need to look further into.

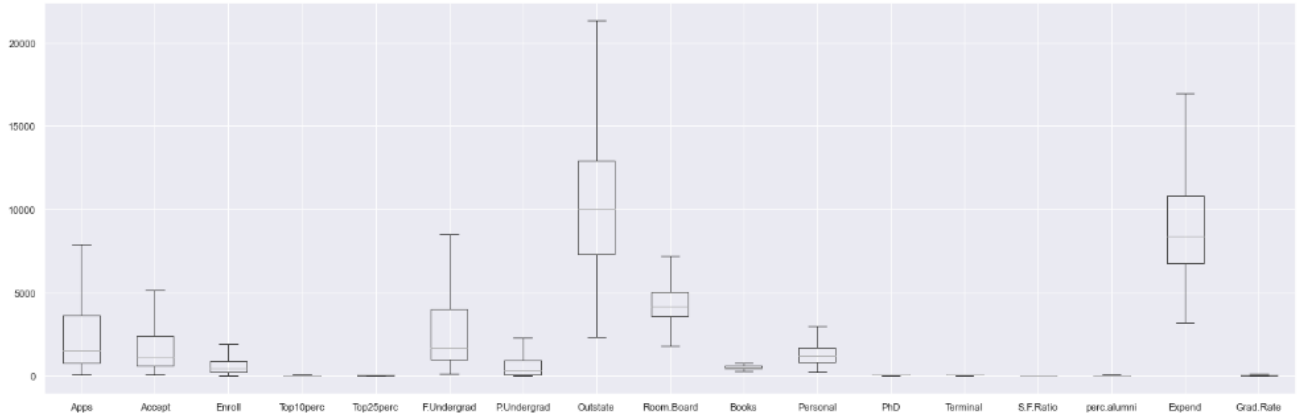
Before outlier treatment :

<AxesSubplot:>



(Fig 6 :- Before treatment)

After outlier treatment :



(Fig 7 :- After treatment)

We can see from above graph that the outliers in the data have been successfully treated as compared to previous case.

If we were to treat the outliers, it is always a good practice to treat the outliers after scaling because, even if we treat the outliers before scaling, there are chances we might encounter additional outliers after scaling which might need further treatment. Thus for comparative analysis, we will be using data treated for outliers after scaling.

2.5 Extract the eigenvalues and eigenvectors. [print both]

Solution:

Covariance matrix helps in seeing the variance in the data; it's having variance stored in them. Eigen values help in explaining the variance and they are calculated based on covariance matrix. By comparing the eigenvalues for above two cases, i.e., with and without treatment of outliers, we can see that a significant amount of variance was not explained by the principal components after treatment of outliers which could be useful in our case study.

Eigen Values of the Covariance Matrix are as follows :

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

(Matrix 3)

Eigen vectors of the Covariance Matrix for corresponding Eigen values are as follows :

```
[[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
   5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
   1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
  -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
  -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
   1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
  -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
  -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
  -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
  -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
  -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
  -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
  -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
 [-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
   2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
   4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
   1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
  -8.23443779e-01]
 [-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
   5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
  -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
  -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
   3.54559731e-01]
 [-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
  -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
   1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
   3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
```

```
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892265e-02
 -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
  1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
  3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
 -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

(Matrix 4)

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

Solution: -

Output:-

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

```
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02 -1.27028371e-01 -5.34724832e-01
 -1.40166326e-01  9.12555212e-02 -1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02 -6.60375454e-02 -5.19443019e-01
 -2.04719730e-01  1.54927646e-01 -2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01 -2.89848401e-01 -1.61189487e-01
  7.93882496e-02  4.87045875e-01  2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01 -1.46989274e-01  1.73142230e-02
  2.16297411e-01 -4.73400144e-02  2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01  2.26743985e-01  7.92734946e-02
 -7.59581203e-02 -2.98118619e-01 -2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01 -2.08064649e-01  2.69129066e-01
  1.09267913e-01  2.16163313e-01  5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

(Matrix 5)

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

Solution: -

Output: -

explicit form of the first PC in terms of the eigenvectors in two places of decimals :

```
array([ 25.,  46.,  64.,  99., 133., 148., 151., 180., 205., 211., 207.,
       239., 271., 253., 274., 306., 331.])
```

Columns of dataset:

```
Index(['Apps', 'Accept', 'Enroll', 'Top10perc', 'Top25perc', 'F.Undergrad',
       'P.Undergrad', 'Outstate', 'Room.Board', 'Books', 'Personal', 'PhD',
       'Terminal', 'S.F.Ratio', 'perc.alumni', 'Expend', 'Grad.Rate'],
      dtype='object')
```

- Explicit form of first PC = (25 * Apps) + (46 * Accept) + (64 * Enroll) + (99 * Top10perc) + (133 * Top25perc) + (148 * F.Undergrad) + (151 * P.Undergrad) + (180 *Outstate)* + *(205 * Room.Board)* + *(211 * Books)* + *(207 * Personal)* + *(239* PhD) + (271 * Terminal) + (253 * S.F.Ratio) + (274 * perc.alumni) + (306 * Expend) + (331 * Grad.Rate)

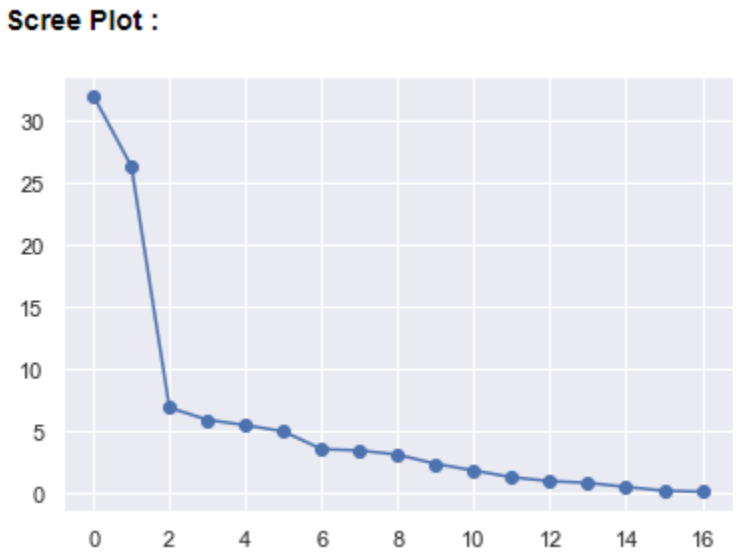
2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

Solution: -

The cumulative values of eigenvalues in ascending order of eigenvalues is as follows

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

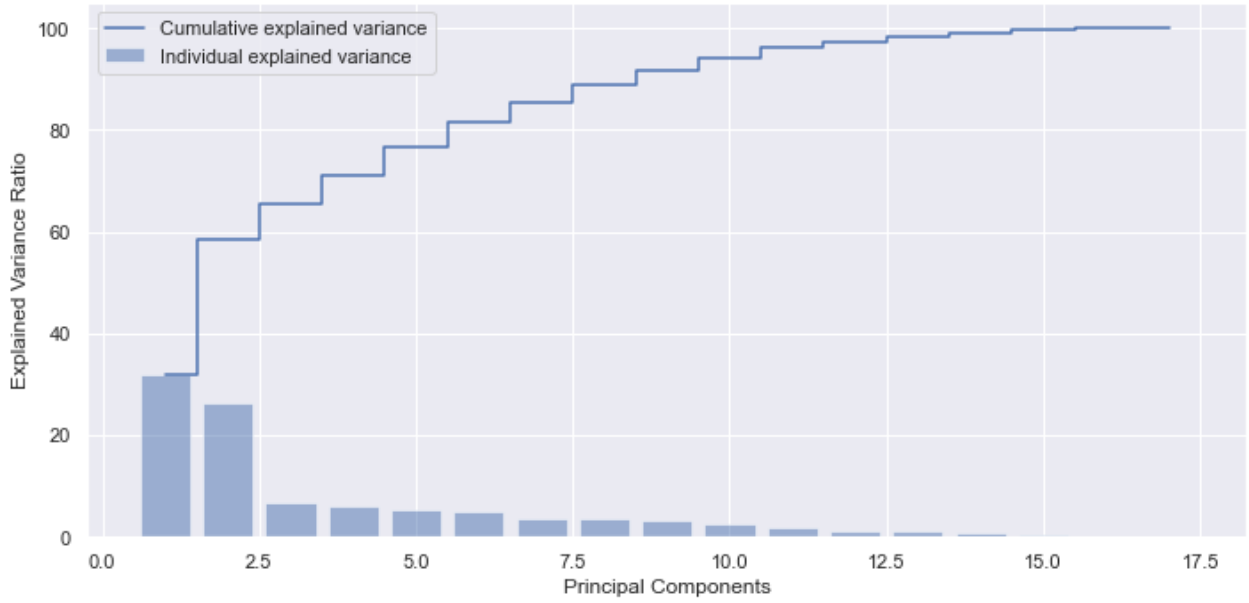
Plotting the scree plot



(Fig 8 plot)

Plot Cumulative explained variance and individual explained variance vs Principal Components

Cumulative and Individual explained variance in a single plot :



(Fig 9 plot)

Visually we can observe that there is a steep drop in variance explained with increase in number of PC's.

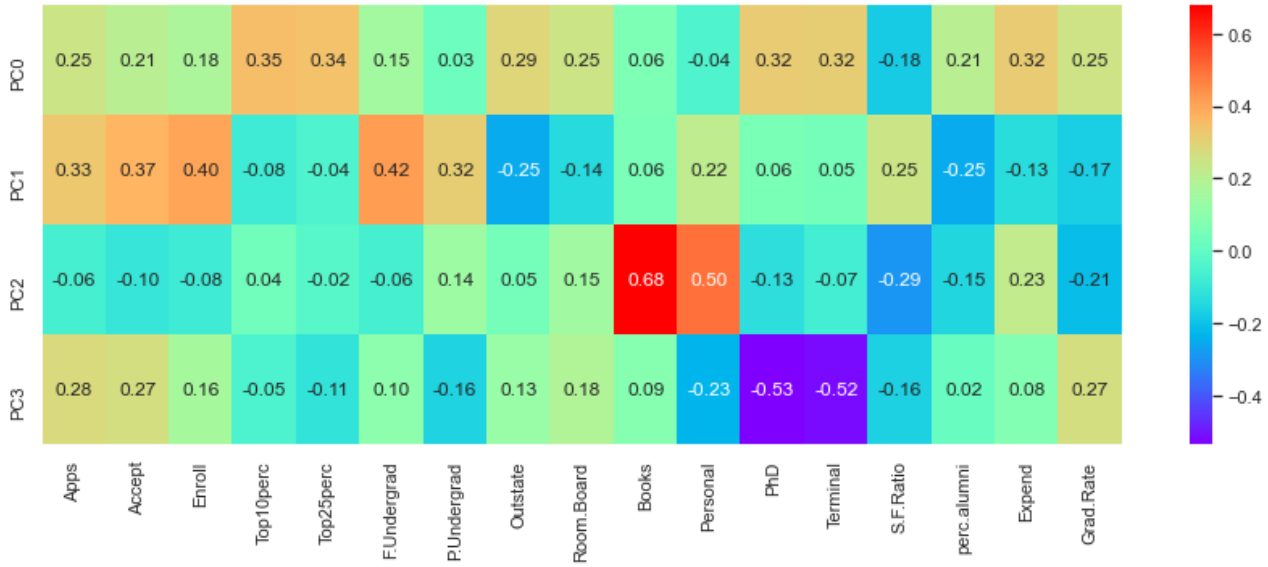
There are a few criteria's which can be considered while selecting the number of principal components:

- 1 eigenvalue: There are 4 PCs with eigenvalues greater than 1 and generally these are considered more significant than others.
- For accurate representation the minimum number of PC's should be $1/4^{th}$ of the total variables, in this case four components, should be taken into consideration. The first four components explain 71.19% of variance in data.
- One could select the top components such that the cumulative variance exceeds a threshold, such as 80%. In this case, 6 components.

In this manner, we are able to cover 85.22% variance in the data with minimum correlation between the components, reducing redundancy. The 7 principal components can then be used in place of the original 17 predictors, reducing dimensionality.

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

Solution: -



(Fig 10 heat map)

Interpretation of the principal components is based on finding which variables are most strongly correlated with each component, i.e., which of these numbers are large in magnitude, the farthest from zero in either direction. And as per that we deduced all the variables are getting covered by the 7 PCs in the following way:

PC0 = Outstate, Expend, Top10perc, Top25perc
PC1 = Apps, Accept, Enroll, F.Undergrad, P.Undergrad, perc.alumni
PC2 = Books
PC3 = PhD, Terminal
PC4 = Room.Board

First Principal Component Analysis – PC0

Variables: Outstate (Number of students for whom the particular college or university is Out-of-state tuition), Expend (The Instructional expenditure per student), Top10perc (Percentage of new students from top 10% of Higher Secondary class), Top25perc (Percentage of new students from top 25% of Higher Secondary class)

It is most positively correlated with these two variables as compared to others. We know that out-of-state tuition is typically more expensive than In-state tuition. Out-of-state students pay more simply because they do not pay taxes to the state in which the university is located. In-state residents, on the other hand, have been supporting the state, and thus indirectly funding the university, all their lives. Thus, lower tuition costs are the state's way of both rewarding its residents for their contributions and accounting for the tax dollars they have already paid to support their state's schools. And from this relation we can say the PC0 correlates with the fact that students for whom the particular college or university is Out-of-state have more expenditure when compared to other students.

Top10perc (Percentage of new students from top 10% of Higher Secondary class), Top25perc (Percentage of new students from top 25% of Higher Secondary class) are also positive significant contributors here, which can further tell us that more students are preferred if they belong to these two categories.

Second Principal Component Analysis – PC1

Variables: Apps (Number of applications received), Accept (Number of applications accepted), Enroll (Number of new students enrolled), F.Undergrad (Number of full-time undergraduate students), perc.alumni (Percentage of alumni who donate)

It is most positively correlated with these variables and slightly negatively but most related with perc.alumni as compared to others. Since there is positive correlation between Apps, Accept, Enroll, and F.Undergrad we can deduce that if the number of applications received is higher, then chances of number of applications accepted are also higher, and from them enrollments of new students are also higher. These new enrolled students are more likely to be full-time undergraduate students. It also shows that alumni are less likely to donate if the college has higher number of students, i.e. Apps, Accept and Enroll since the alumni might think the institution makes enough money from the higher number of students.

Third Principal Component Analysis – PC2

Variables: Books (Estimated book costs for a student)
This PC is closely positively correlated with books and also Personal (Estimated personal spending for a student) (not as much as PC6 but still a significant amount but makes sense here). It shows that books might constitute of a significant part of personal spending of students.

Fourth Principal Component Analysis – PC3

Variables: PhD (Percentage of faculties with Ph.D.'s), Terminal (Percentage of faculties with terminal degree) This PC is highly negatively correlated with PhD and Terminal. But the values of them are same. It can imply that the faculties with PhD indeed have PhD as their terminal degree, i.e., the highest academic degree that can be awarded in a particular field. It can be a measure of faculties with highest level of qualification in their fields and there is a high chance that it is PhD.

Fifth Principal Component Analysis – PC4

Variables: Room.Board (Cost of Room and board)
It is most positively correlated with Room.Board. This component can be a measure of the Cost of Room and board for an institution.

Thankyou….