Projects on Advanced Statistics

Project Report by: Mr.J.B.Ashtekar (PGPDSBA group 9)

Mail Id:ashtekarjaydeep@yahoo.in

Contact:9689630190

Date of submission:12/01/2020

Content

Project on ANOVA

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

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

Ans:

1. For variable A

- null hypothesis is $\mu1=\mu2=\mu3$
- alternate hypothesis is atleast one out of $\mu1,\mu2$ or $\mu3$ is not same.

2. For variable B

- null hypothesis is $\mu1=\mu2=\mu3$
- alternate hypothesis is at least one out of $\mu1,\mu2$ or $\mu3$ is not same.

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

Ans:

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

For A as P value is less than 0.05(4.578242e-07) so null will be rejected(means A has significant effect on relief)

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

Ans:

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

For B as P value is less than 0.05(0.00135) so null will be rejected.(means B has significant effect on relief)

1.4) Analyse the effects of one variable on another with the help of an interaction plot. What is an interaction between two treatments?

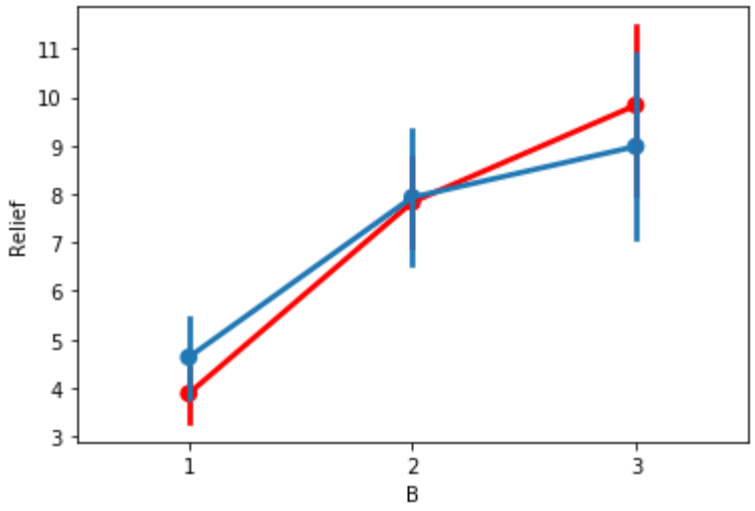
Ans:so from this interaction plot it can be inferred that

1.At level 1 B gives good relief than A

2.At level 2 relief hours are almost same.

3. At level 3 A gives good relief than B.

4.Rate of increase of the relief is more for than B.



A:red,B:blue

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

Ans:

```
Table1:
Effect of A and B=

             df   sum_sq      mean_sq           F         PR(>F)
```

```
C(A)       2.0  220.02  110.010000  109.832850  8.514029e-15
C(B)       2.0  123.66   61.830000   61.730435  1.546749e-11
Residual  31.0   31.05    1.001613         NaN           NaN
```

Table2:
Effect of A and B(interaction with Volunteer)=

```
                     df      sum_sq     mean_sq          F        PR(>F)
C(A)                2.0  220.020000  110.010000  78.971721  9.123327e-11
C(B):C(Volunteer)  11.0  124.063333   11.278485   8.096367  1.863192e-05
Residual           22.0   30.646667    1.393030        NaN           NaN
```

Table3:
Effect of B and A(interaction with Volunteer)=

```
                     df      sum_sq    mean_sq          F        PR(>F)
C(B)                2.0  123.660000  61.830000  44.183413  1.974356e-08
C(A):C(Volunteer)  11.0  220.283333  20.025758  14.310307  1.370125e-07
Residual           22.0   30.786667   1.399394        NaN           NaN
```

Table4:
Effect of A(interaction with Volunteer)and B(interaction with Volunteer)=

```
                     df      sum_sq    mean_sq          F    PR(>F)
C(B):C(Volunteer)  11.0  124.063333  11.278485   5.925216  0.000779
C(A):C(Volunteer)   8.0  220.211111  27.526389  14.461146  0.000006
Residual           16.0   30.455556   1.903472        NaN       NaN
```

Table5:
Effect of (A + B)(interaction with Volunteer)=

```
                     df      sum_sq    mean_sq          F    PR(>F)
C(B):C(Volunteer)  11.0  124.063333  11.278485   5.925216  0.000779
C(A):C(Volunteer)   8.0  220.211111  27.526389  14.461146  0.000006
Residual           16.0   30.455556   1.903472        NaN       NaN
```

Table6:
Effect of A(interaction with Volunteer)=

```
                     df      sum_sq    mean_sq         F    PR(>F)
C(A):C(Volunteer)  11.0  220.283333  20.025758  3.111872  0.009697
Residual           24.0  154.446667   6.435278       NaN       NaN
```

Table7:
Effect of B(interaction with Volunteer)=

```
                     df      sum_sq    mean_sq         F    PR(>F)
C(B):C(Volunteer)  11.0  124.063333  11.278485  1.079855  0.416083
Residual           24.0  250.666667  10.444444       NaN       NaN
```

Table8:
Effect of B(interaction with Volunteer)=

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

Table9:
Effect of B(interaction with Volunteer)=

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

- All above combinations has significant effect on relief except B(interaction with Volunteer)(Table 7)
- It has p value 0.41608(greater than 0.05)

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

Ans:Business Implications:

- From this ANOVA analysis we conclude that

- Individual A and B has effect on relief.But combination of A and B increses the relief hours.
- A is more significant than B in the compound AB.

Project on PCA

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

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

Ans:

A.Univariate Analysis:

1. CSV file and various libraries are imported.

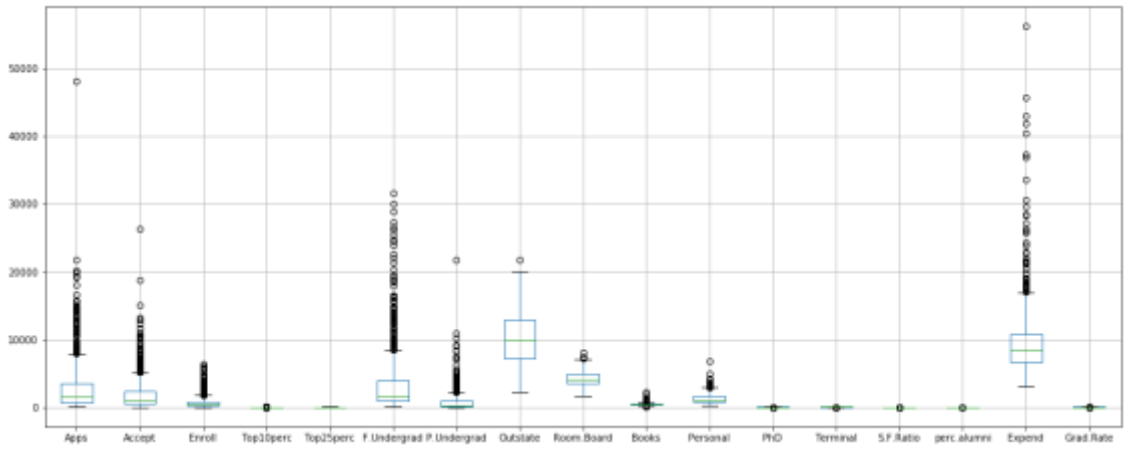
2. Description of each variable is obtained by describe function.

3. Size of the dataset is 777 rows with 17 columns (attributes).

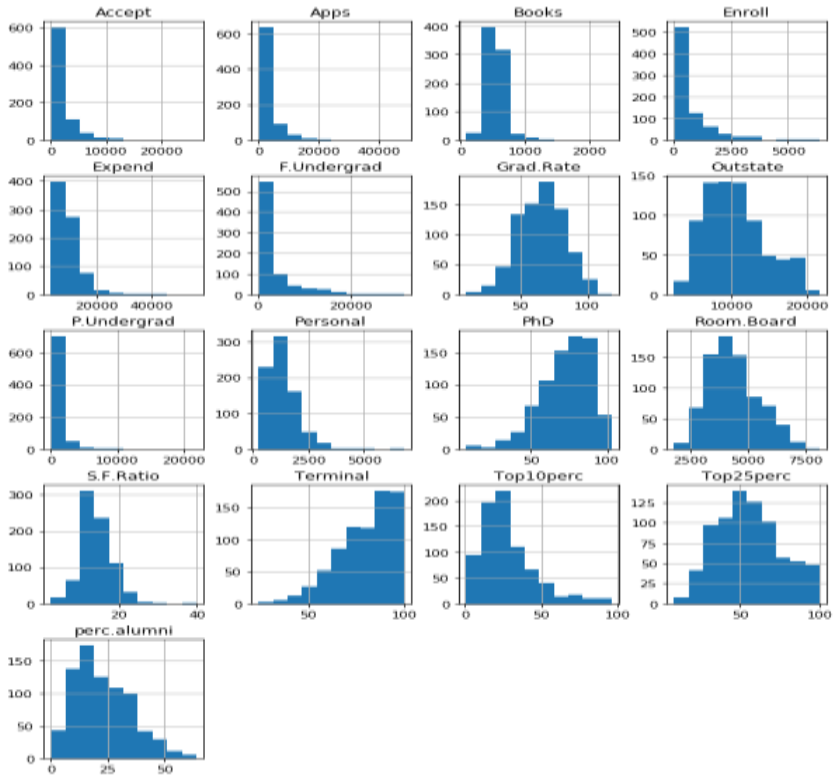
4. All are numeric variable except name.

5. No null values in dataset.

6. Each attribute has outliers except top25per.



7. Data is highly skewed. Accept, Apps, Expend, Enroll, F.undergrad, SF ratio, top10perc, personal these are the right skewed attributes. Phd and terminal are left skewed and remaining top25perc are symmetric about mean.



8. For all attributes Mean, mode and median are calculated.

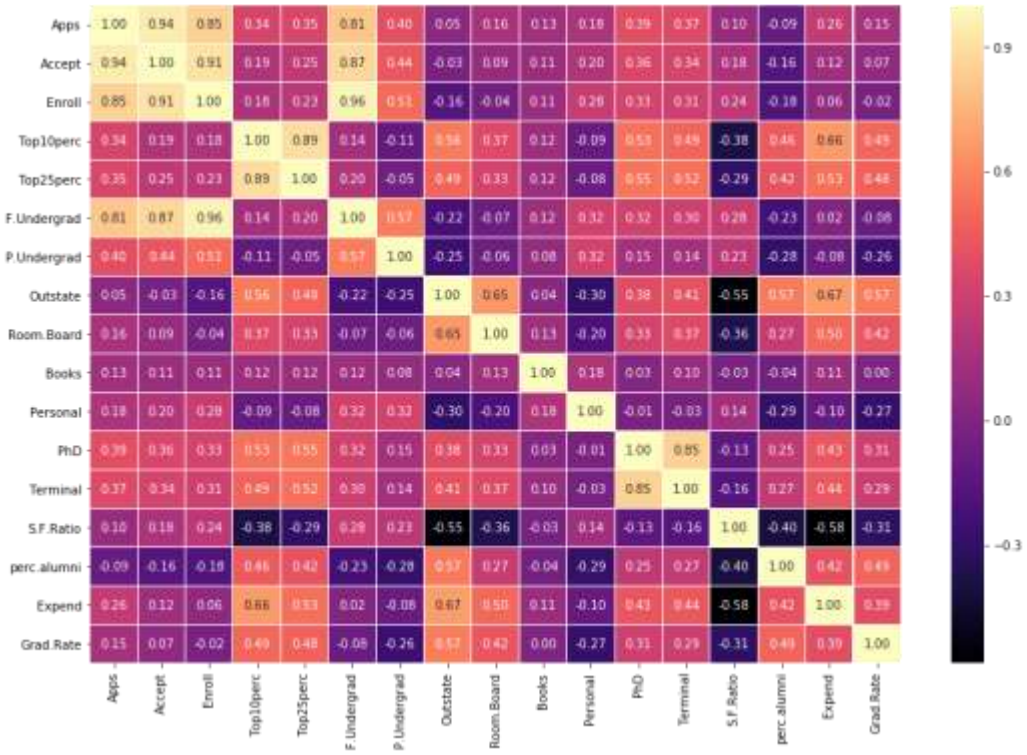
It is observed that data is **Multimodal.**

Expend has maximum modes 33.

Apps 3
Accept 1
Enroll 2
Top10perc 1
Top25perc 2
F.Undergrad 7
P.Undergrad 1
Outstate 1
Room.Board 1
Books 1
Personal 1
PhD 1
Terminal 1
S.F.Ratio 1
perc.alumni 1
Expend 33
Grad.Rate 1

B) Multivariate Analysis:

1.Correlation:

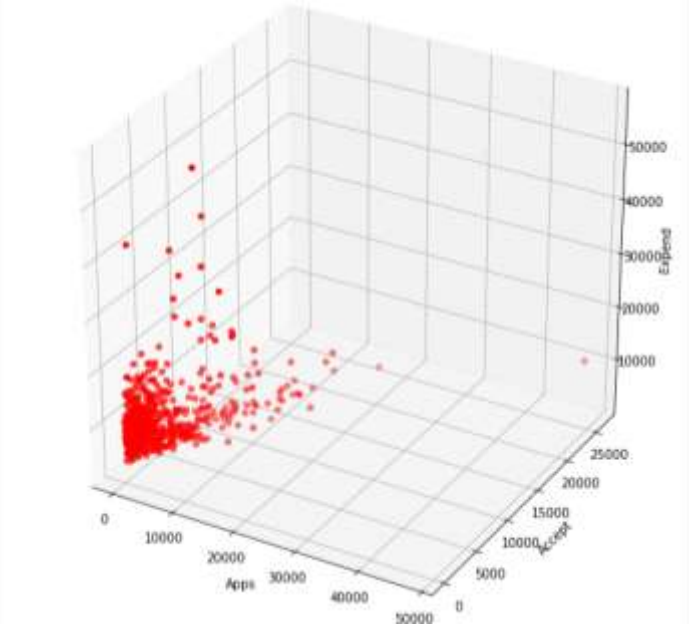


Following is the rows and columns which has good correlation with each other.

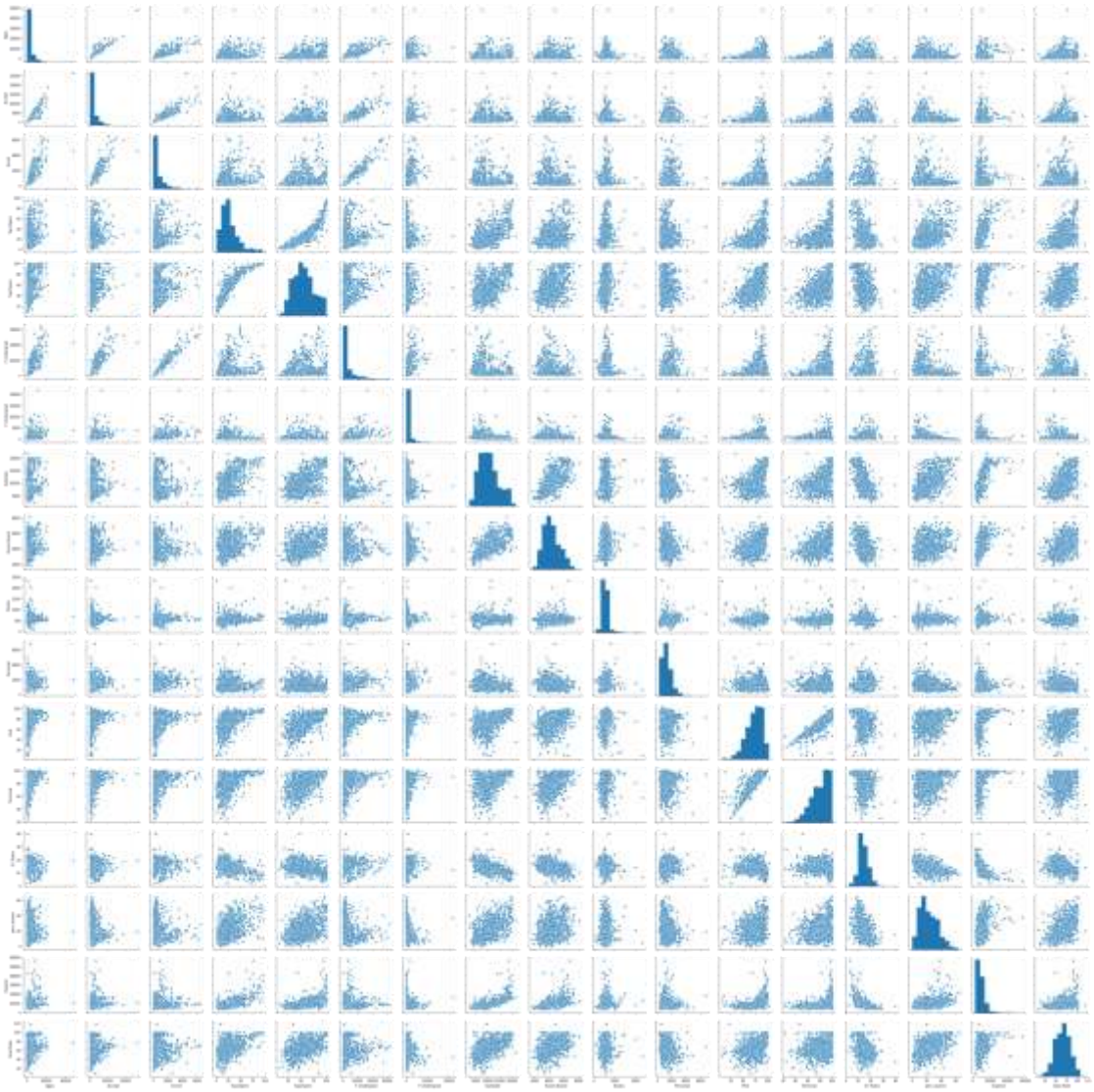
[edu.corr()[(edu.corr()>0.8)]].*Rows and columns are deleted in Excel in following table.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	PhD	Terminal
Apps		0.943451	0.846822			0.814491		
Accept	0.943451		0.911637			0.874223		
Enroll	0.846822	0.911637				0.96464		
Top10perc					0.891995			
Top25perc				0.891995				
F.Undergrad	0.814491	0.874223	0.96464					
PhD								0.849587
Terminal							0.849587	

2. 3D scatter plot (for Apps, Accept and Expend which shows good correlation at initial part.)



3. Pairplot: Diagonal shows histogram which is showing skewness already explained above. Similarly correlation exist as shown in above table.

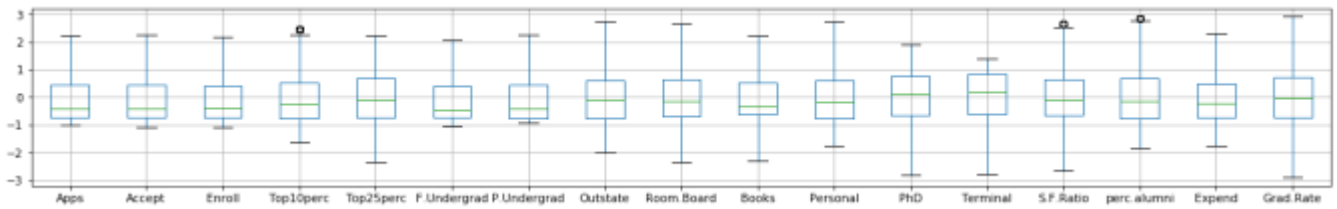


2.2) Scale the variables and write the inference for using the type of scaling function for this case study.

Ans:

1. Scaling is needed because some variables are in terms of ratio, percentages, expenses, numbers. Scaling converts all attributes in one scale.

2. Scaling used is $(x-\mu)/\sigma$ which converted all variables on scale of 3 to -3.



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

Ans:Following table shows the difference between covariance matrix and correlation matrix.It shows great difference.As covariance shows the variance from each other.Correlation shows the linear relationships between each other.

"Covariance" gives the direction of the linear relationship between attributes. "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. You can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.(as per definitions)

But if values are standardized then it shows same behavior.

	Apps	**Accept**	**Enroll**	**Top 10perc**	**Top 25perc**	**F.Undergrad**	**P.Undergrad**	**Outstate**	**Room.Board**	**Books**	**Personal**	**PhD**	**Terminal**	**S.F. Ratio**	**perc.alumni**	**Expend**	**Grad.Rate**
Apps	0.00	0.01	0.05	-0.02	0.01	0.05	0.12	0.02	0.02	0.10	0.05	0.07	0.07	0.03	-0.01	-0.02	0.00
Accept	0.01	0.00	0.02	0.03	0.03	0.02	0.13	0.02	0.03	0.10	0.06	0.07	0.07	0.01	-0.01	0.04	0.01
Enroll	0.05	0.02	0.00	-0.01	0.00	0.00	0.13	0.00	0.02	0.09	0.06	0.05	0.05	0.04	-0.04	-0.01	0.00
Top 10perc	-0.02	0.03	-0.01	0.00	0.02	-0.03	-0.07	0.00	-0.01	0.03	-0.02	0.01	0.02	0.00	0.00	0.00	0.00
Top 25perc	0.01	0.03	0.00	0.02	0.00	-0.02	-0.05	0.00	0.00	0.05	-0.01	0.01	0.00	0.00	0.00	0.05	0.00
F.Undergrad	0.05	0.02	0.00	-0.03	-0.02	0.00	0.13	-0.01	0.01	0.09	0.04	0.04	0.04	0.05	-0.06	-0.02	0.00
P.Undergrad	0.12	0.13	0.13	-0.07	-0.05	0.13	0.00	-0.10	-0.01	0.04	0.02	-0.02	-0.02	0.14	-0.14	-0.12	-0.01
Outstate	0.02	0.02	0.00	0.00	0.00	-0.01	-0.10	0.00	0.00	-0.03	-0.03	0.01	0.01	-0.02	0.00	0.10	0.00
Room.Board	0.02	0.03	0.02	-0.01	0.00	0.01	-0.01	0.00	0.00	-0.02	-0.02	0.01	0.01	-0.01	0.00	0.08	0.00
Books	0.10	0.10	0.09	0.03	0.05	0.09	0.04	-0.03	-0.02	0.00	0.06	0.11	0.06	0.02	0.00	0.04	-0.01
Personal	0.05	0.06	0.06	-0.02	-0.01	0.04	0.02	-0.03	-0.02	0.06	0.00	0.00	0.00	0.04	-0.02	-0.07	-0.02
PhD	0.07	0.07	0.05	0.01	0.01	0.04	-0.02	0.01	0.01	0.11	0.00	0.00	0.01	0.00	0.00	0.08	0.01
Terminal	0.07	0.07	0.05	0.02	0.00	0.04	-0.02	0.01	0.01	0.06	0.00	0.01	0.00	0.01	0.00	0.09	0.00

S.F. Ratio	0.03	0.01	0.04	0.00	0.00	0.05	0.14	-0.02	-0.01	0.02	0.04	0.00	0.01	0.00	-0.01	-0.07	0.00
perc.alumni	-0.01	-0.01	-0.04	0.00	0.00	-0.06	-0.14	0.00	0.00	0.00	-0.02	0.00	0.00	-0.01	0.00	0.05	0.00
Expend	-0.02	0.04	-0.01	0.00	0.05	-0.02	-0.12	0.10	0.08	0.04	-0.07	0.08	0.09	-0.07	0.05	0.00	0.03
Grad.Rate	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	-0.01	-0.02	0.01	0.00	0.00	0.00	0.03	0.00

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

Ans:

1. Before scaling outliers are 584 and after scaling outliers (abs(z)>3) are 163.

Apps 70	21
Accept 73	22
Enroll 79	23
Top10perc 39	11
Top25perc 0	0
F.Undergrad 97	24
P.Undergrad 67	23
Outstate 1	0
Room.Board 7	0
Books 46	11
Personal 20	9
PhD 8	2
Terminal 8	2
S.F.Ratio 12	2
perc.alumni 5	0
Expend 48	13
Grad.Rate 4	0
584 before scaling	163 after scaling

2. Then outliers are imputed by median then dataset has no outliers then converted to z-score which gave some outliers which are ignored then.

2.5) Build the covariance matrix, eigen values and eigenvector.

Ans:

1.Covariance matrix

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.001	0.957	0.898	0.322	0.365	0.862	0.520	0.065	0.188	0.236	0.230	0.465	0.435	0.127	-0.101	0.243	0.151
Accept	0.957	1.001	0.936	0.224	0.274	0.898	0.573	-0.005	0.120	0.209	0.257	0.428	0.404	0.189	-0.166	0.162	0.079
Enroll	0.898	0.936	1.001	0.172	0.231	0.969	0.642	-0.156	-0.024	0.202	0.340	0.382	0.355	0.275	-0.223	0.054	-0.023
Top10perc	0.322	0.224	0.172	1.001	0.915	0.111	-0.180	0.563	0.358	0.154	-0.117	0.545	0.507	-0.388	0.456	0.658	0.494
Top25perc	0.365	0.274	0.231	0.915	1.001	0.181	-0.099	0.490	0.331	0.170	-0.087	0.552	0.528	-0.298	0.417	0.574	0.480
F.Undergrad	0.862	0.898	0.969	0.111	0.181	1.001	0.697	-0.226	-0.055	0.208	0.360	0.362	0.335	0.325	-0.286	0.000	-0.082
P.Undergrad	0.520	0.573	0.642	-0.180	-0.099	0.697	1.001	-0.355	-0.068	0.123	0.344	0.128	0.122	0.371	-0.420	-0.202	-0.265
Outstate	0.065	-0.005	-0.156	0.563	0.490	-0.226	-0.355	1.001	0.656	0.005	-0.326	0.392	0.413	-0.574	0.566	0.776	0.573
Room.Board	0.188	0.120	-0.024	0.358	0.331	-0.055	-0.068	0.656	1.001	0.109	-0.220	0.342	0.380	-0.377	0.273	0.581	0.426
Books	0.236	0.209	0.202	0.154	0.170	0.208	0.123	0.005	0.109	1.001	0.240	0.137	0.160	-0.009	-0.043	0.150	-0.008
Personal	0.230	0.257	0.340	-0.117	-0.087	0.360	0.344	-0.326	-0.220	0.240	1.001	-0.012	-0.032	0.174	-0.306	-0.163	-0.291
PhD	0.465	0.428	0.382	0.545	0.552	0.362	0.128	0.392	0.342	0.137	-0.012	1.001	0.864	-0.130	0.249	0.511	0.310
Terminal	0.435	0.404	0.355	0.507	0.528	0.335	0.122	0.413	0.380	0.160	-0.032	0.864	1.001	-0.151	0.266	0.525	0.293
S.F.Ratio	0.127	0.189	0.275	-0.388	-0.298	0.325	0.371	-0.574	-0.377	-0.009	0.174	-0.130	-0.151	1.001	-0.413	-0.655	-0.309
perc.alumni	-0.101	-0.166	-0.223	0.456	0.417	-0.286	-0.420	0.566	0.273	-0.043	-0.306	0.249	0.266	-0.413	1.001	0.464	0.492

Expend	0.243	0.162	0.054	0.658	0.574	0.000	-0.202	0.776	0.581	0.150	-0.163	0.511	0.525	-0.655	0.464	1.001	0.416
Grad.Rate	0.151	0.079	-0.023	0.494	0.480	-0.082	-0.265	0.573	0.426	-0.008	-0.291	0.310	0.293	-0.309	0.492	0.416	1.001

2.Eigen Vectors

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

```
 -8.14247697e-02  5.60392799e-01  6.72405494e-02 -2.11561014e-01
 -1.59275617e-03]
[-2.44186588e-01 -1.31909124e-01 -2.11379165e-02 -5.80894132e-01
  6.91080879e-02  2.37267409e-01  9.45210745e-02  3.89639465e-01
 -2.20526518e-01 -2.55107620e-01 -4.75297296e-01  9.85725168e-03
  2.67779296e-02 -1.07365653e-01  1.77715010e-02 -1.00935084e-01
 -2.82578388e-02]
[-9.67082754e-02  9.39739472e-02 -6.97121128e-01  3.61562884e-02
 -3.54056654e-02  6.38604997e-01 -1.11193334e-01 -2.39817267e-01
  2.10246624e-02  9.11624912e-02  4.35697999e-02  4.36086500e-03
  1.04624246e-02  5.16224550e-02  3.54343707e-02 -2.86384228e-02
 -8.06259380e-03]
[ 3.52299594e-02  2.32439594e-01 -5.30972806e-01  1.14982973e-01
  4.75358244e-04 -3.81495854e-01  6.39418106e-01  2.77206569e-01
  1.73715184e-02 -1.27647512e-01  1.51627393e-02 -1.08725257e-02
  4.54572099e-03  9.39409228e-03 -1.18604404e-02  3.38197909e-02
  1.42590097e-03]
[-3.26410696e-01  5.51390195e-02  8.11134044e-02  1.47260891e-01
  5.50786546e-01  3.34444832e-03  8.92320786e-02 -3.42628480e-02
  1.66510079e-01  1.00975002e-01 -3.91865961e-02  1.33146759e-02
  1.25137966e-02 -7.16590441e-02  7.02656469e-01 -6.38096394e-02
  8.31471932e-02]
[-3.23115980e-01  4.30332048e-02  5.89785929e-02  8.90079921e-02
  5.90407136e-01  3.54121294e-02  9.16985445e-02 -9.03076644e-02
  1.12609034e-01  8.60363025e-02 -8.48575651e-02  7.38135022e-03
 -1.79275275e-02  1.63820871e-01 -6.62488717e-01  9.85019644e-02
 -1.13374007e-01]
[ 1.63151642e-01  2.59804556e-01  2.74150657e-01  2.59486122e-01
  1.42842546e-01  4.68752604e-01  1.52864837e-01  2.42807562e-01
 -1.53685343e-01 -4.70527925e-01  3.63042716e-01  8.85797314e-03
  1.83059753e-02 -2.39902591e-01 -4.79006197e-02  6.19970446e-02
  3.83160891e-03]
[-1.86610828e-01 -2.57092552e-01  1.03715887e-01  2.23982467e-01
 -1.28215768e-01  1.25669415e-02  3.91400512e-01 -5.66073056e-01
 -5.39235753e-01 -1.47628917e-01 -1.73918533e-01 -2.40534190e-02
 -8.03169296e-05 -4.89753356e-02  3.58875507e-02  2.80805469e-02
 -7.32598621e-03]
[-3.28955847e-01 -1.60008951e-01 -1.84205687e-01 -2.13756140e-01
  2.24240837e-02 -2.31562325e-01 -1.50501305e-01 -1.18823549e-01
  2.42371616e-02 -8.04154875e-02  3.93722676e-01  1.05658769e-02
  5.60069250e-02 -6.90417042e-01 -1.26667522e-01  1.28739213e-01
  1.45099786e-01]
[-2.38822447e-01 -1.67523664e-01  2.45335837e-01  3.61915064e-02
 -3.56843227e-01  3.13556243e-01  4.68641965e-01  1.80458508e-01
  3.15812873e-01  4.88415259e-01  8.72638706e-02 -2.51028410e-03
  1.48410810e-02 -1.59332164e-01 -6.30737002e-02 -7.09643331e-03
```

-3.29024228e-03]]

3.Eigen Values

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

Eigen vectors are basically directions and eigen value is magnitude. Eigen value explains variance in new feature axis. The more directions you have along which you understand the behaviour of a linear transformation, the easier it is to understand the linear transformation; so you want to have as many linearly independent eigenvectors as possible associated to a single linear transformation.

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

The equation for first PC would be the cross multiplication of variables and the Eigen vectors of 0 index.

Eigen vector for PC1:

array([[2.62171542e-01, 2.30562461e-01, 1.89276397e-01,
3.38874521e-01, 3.34690532e-01, 1.63293010e-01,
2.24797091e-02, 2.83547285e-01, 2.44186588e-01,
9.67082754e-02, -3.52299594e-02, 3.26410696e-01,
3.23115980e-01, -1.63151642e-01, 1.86610828e-01,
3.28955847e-01, 2.38822447e-01]]))

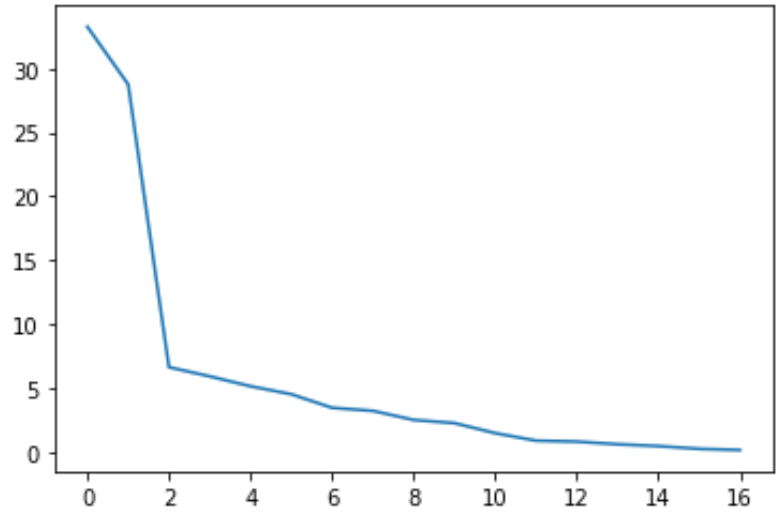
PC1 equation:

0.262172**Apps*+0.230562*Accept+0.189276**Enroll*+0.338875*Top10perc+0.334691*T*
op25perc+0.163293F.Undergrad+0.02248*P.Undergrad*+0.283547*Outstate+0.244187*
Room.Board+0.096708 Books-0.03523 *Personal*+0.326411 PhD+0.323116 *Terminal*-
0.163152* S.F.Ratio+0.186611**perc.alumni*+0.328956 * *Expend* +
0.238822*Grad.Rate

2.7) Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	0.262172	0.230562	0.189276	0.338875	0.334691	0.163293	0.02248	0.283547	0.244187	0.096708	-0.03523	0.326411	0.323116	-0.16315	0.186611	0.328956	0.238822
1	0.314136	0.344624	0.382813	-0.09932	-0.05951	0.398636	0.35755	-0.25186	-0.13191	0.093974	0.23244	0.055139	0.043033	0.259805	-0.25709	-0.16001	-0.16752
2	-0.08102	-0.10766	-0.08553	0.078829	0.050794	-0.07371	-0.04036	-0.01494	0.021138	0.697121	0.530973	-0.08111	-0.05898	-0.27415	-0.10372	0.184206	-0.24534
3	0.098776	0.11814	0.009307	-0.36912	-0.41682	0.01395	0.225351	0.262975	0.580894	-0.03616	-0.11498	-0.14726	-0.08901	-0.25949	-0.22398	0.213756	-0.03619
4	0.219898	0.189635	0.162315	0.157211	0.144449	0.102728	-0.09568	0.037275	-0.06911	0.035406	-0.00048	-0.55079	-0.59041	-0.14284	0.128216	-0.02242	0.356843
5	0.002188	-0.01652	-0.06808	-0.08887	-0.02763	-0.05165	-0.02454	-0.02039	0.237267	0.638605	-0.3815	0.003344	0.035412	0.468753	0.012567	-0.23156	0.313556
6	-0.02837	-0.01296	-0.01524	-0.25746	-0.23904	-0.03118	-0.01001	0.094537	0.094521	-0.11119	0.639418	0.089232	0.091699	0.152865	0.391401	-0.1505	0.468642
7	-0.08995	-0.13761	-0.14422	0.289539	0.345644	-0.10875	0.123842	0.011272	0.389639	-0.23982	0.277207	-0.03426	-0.09031	0.242808	-0.56607	-0.11882	0.180459
8	-0.13057	-0.14228	-0.05087	0.122468	0.193936	-0.00146	0.634774	0.008366	0.220527	-0.02103	-0.01737	-0.16651	-0.11261	0.153685	0.539236	-0.02424	-0.31581
9	-0.15646	-0.14921	-0.0649	-0.03588	0.006418	-0.00016	0.546346	-0.2318	-0.25511	0.091162	-0.12765	0.100975	0.086036	-0.47053	-0.14763	-0.08042	0.488415
10	-0.08621	-0.04259	-0.04384	0.001778	-0.10213	-0.035	0.252107	0.593433	-0.4753	0.04357	0.015163	-0.03919	-0.08486	0.363043	-0.17392	0.393723	0.087264
11	-0.08998	-0.15886	0.035399	0.039228	-0.14562	0.133556	-0.05025	-0.56039	0.107366	-0.05162	-0.00939	0.071659	-0.16382	0.239903	0.048975	0.690417	0.159332
12	-0.08887	-0.0438	0.061924	-0.06996	0.097028	0.087175	-0.04455	-0.06724	-0.01777	-0.03543	0.01186	-0.70266	0.662489	0.047901	-0.03589	0.126668	0.063074

The variance explained by all the principal components will be [33.26608366671336,
28.755345008170774,
6.617163554717709,
5.898143957623823,
5.123892672339139,
4.498638671547009,
3.4362426556658137,
3.198847173205204,
2.4881075492912657,
2.238396454242057,
1.4511567777537842,
0.8651434488112959,
0.7892466165436466,
0.5806273152471961,
0.43878768621189934,
0.22261871681452408,
0.1315580751014857]



The first PC explain 33.27% of variance in the dataset followed by 28.76% of variance explained by PC2 and so on.

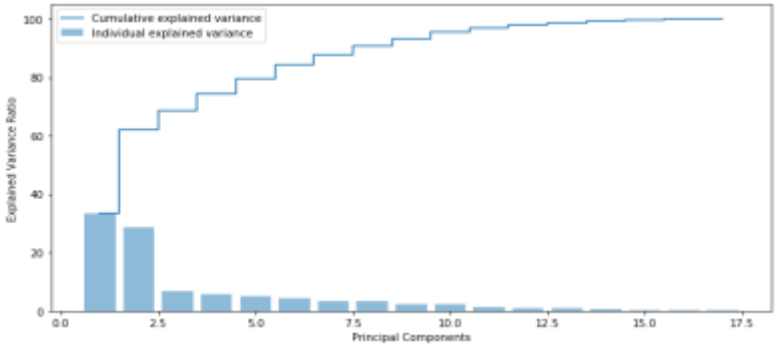
The cumulative variance explained by all the variables as follows:

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

The Cumulative Variance Explained % gives the percentage of variance carried for by the n components altogether. For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components.

In the above array we see that the first feature explains 33.27% of the variance within our data set while the first two explain 58.3 and so on. If we employ 13 features we capture 98.62% of the variance within the dataset, thus we gain very little by implementing an additional feature.

If we are selecting 90% variation captured the PCs will be 8 and so on.So how much cum sum we have to consider it selects the number of PCs.



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

Ans:

This method basically gives the dimensional reductions without losing the variability.So above equation of PC1 is used as a mathematical model for the prediction.