Advanced Statistics

Project Report

Monika Nanda

1/12/2020

Table of Contents

Project Objective

The objective of the report is to explore the project data set in Python and answer the assignment questions by providing insights about the data set.

1 Problem 1

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments.

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

Ingredient A

Null Hypothesis: Average Relief hours are same for all three levels of Ingredient A in the compound

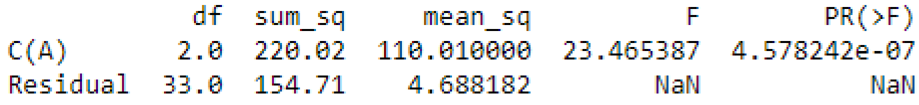
Alternate Hypothesis: Average Relief hours are different for at least one level of Ingredient A in the compound

Ingredient B

Null Hypothesis: Average Relief hours are same for all three levels of Ingredient B in the compound

Alternate Hypothesis: Average Relief hours are different for at least one level of Ingredient B in the compound

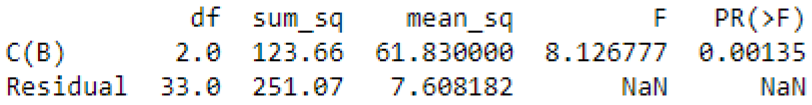
1.2 Perform one-way ANOVA for variable ‘A’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.



```
            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
```

Since the p value is less than the significance level (0.05), we can reject the null hypothesis and state that there is a difference in the mean relief hours of different levels of Ingredient A

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.

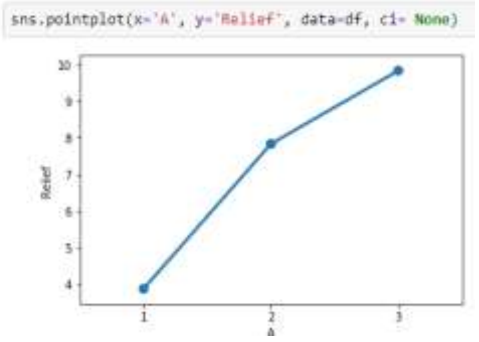


```
            df  sum_sq    mean_sq         F   PR(>F)
C(B)       2.0  123.66  61.830000  8.126777  0.00135
Residual  33.0  251.07   7.608182       NaN      NaN
```

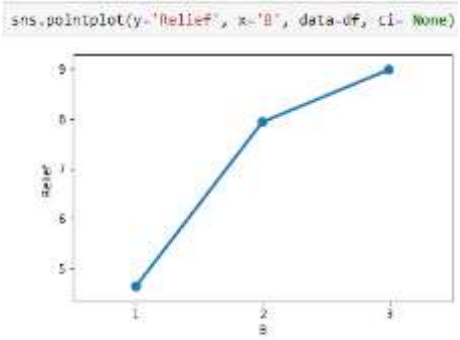
Since the p value is less than the significance level (0.05), we can reject the null hypothesis and state that there is a difference in the mean relief hours of different levels of Ingredient B

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

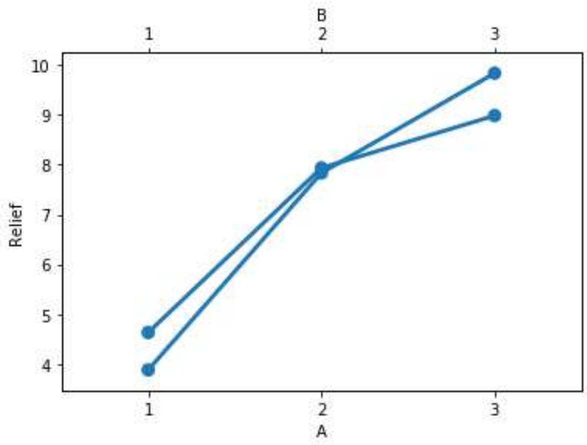
What is an interaction between two treatments?



```
sns.pointplot(x='A', y='Relief', data=df, ci= None)
```

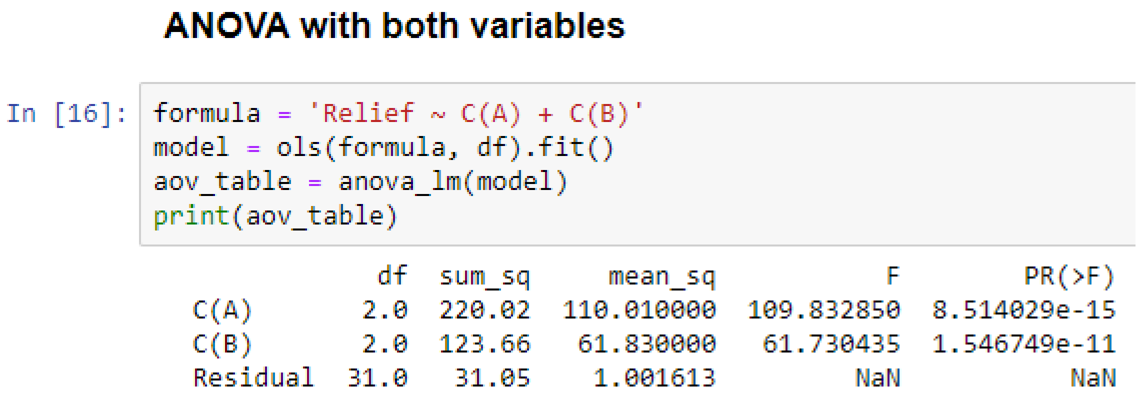


```
sns.pointplot(y='Relief', x='B', data=df, ci= None)
```



From the above graph we can see that level 1 and 3 of Ingredient A are more effective than that of Ingredient B but level 2 of both Ingredients have same effect.There is some but not a defined relationship between the two ingredients A and B.

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

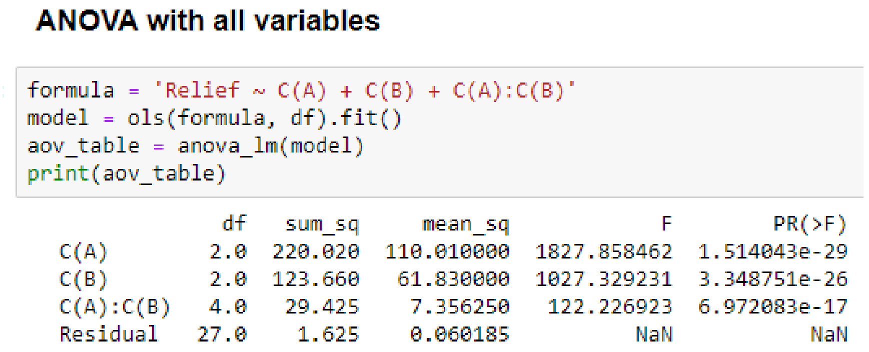


ANOVA with both variables

```
In [16]: formula = 'Relief ~ C(A) + C(B)'
         model = ols(formula, df).fit()
         aov_table = anova_lm(model)
         print(aov_table)

                 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
```

We reject the Null Hypothesis and conclude that the means are different for different values of both variables



ANOVA with all variables

```
formula = 'Relief ~ C(A) + C(B) + C(A):C(B)'
model = ols(formula, df).fit()
aov_table = anova_lm(model)
print(aov_table)

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

We reject all three Null Hypothesis and conclude that the means are different for different values for both variables and there is an interaction between Ingredient A and B.

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

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

- From ANOVA One way Test, we can infer that all levels do not have equal implication on relief hours for both Ingredients A&B

- From Scree Plot, we can see that Level 1 is most effective for both Ingredients A & B
- It can also be observed that level 3 is the least effective for both ingredients A & B as the relief hours are more.
- From the experiment, we can conclude that to minimize the relief hours (i.e. to provide faster relief to severe cases of hay fever), the new compound should be developed using level 1 of Ingredient A and B

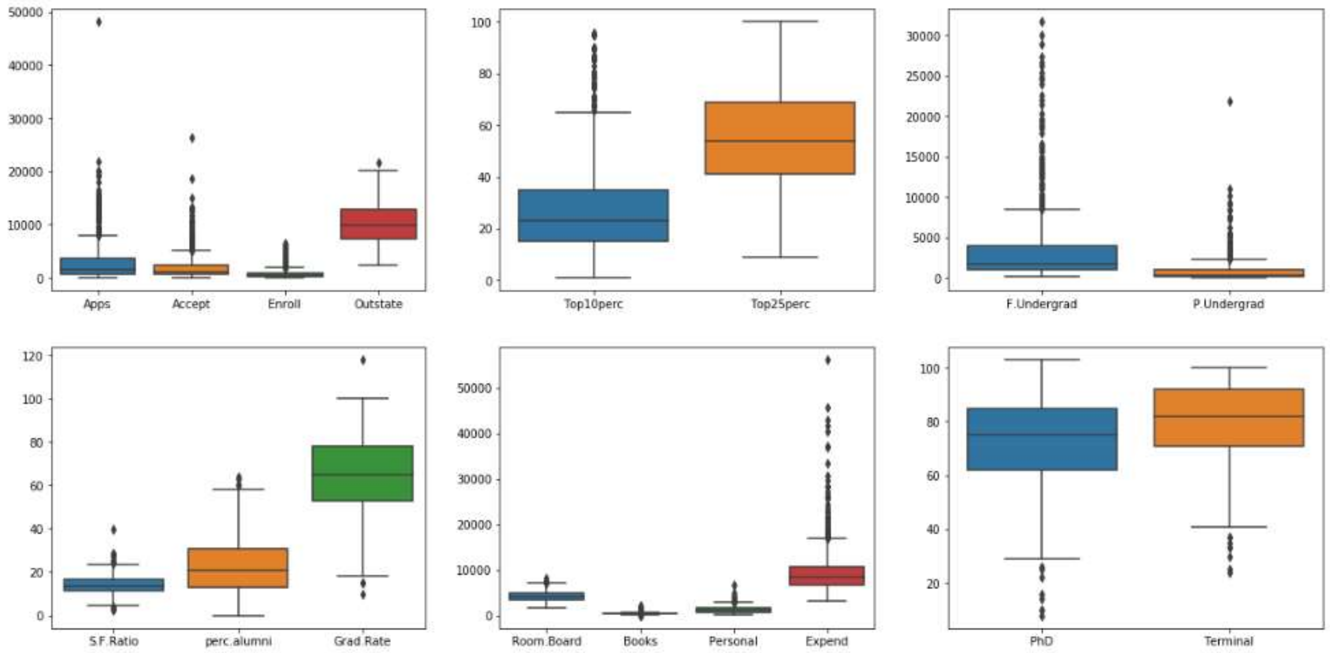
2 Problem 2

The dataset Education - Post 12th Standard.csv is a dataset which contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric.

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

Univariate Analysis

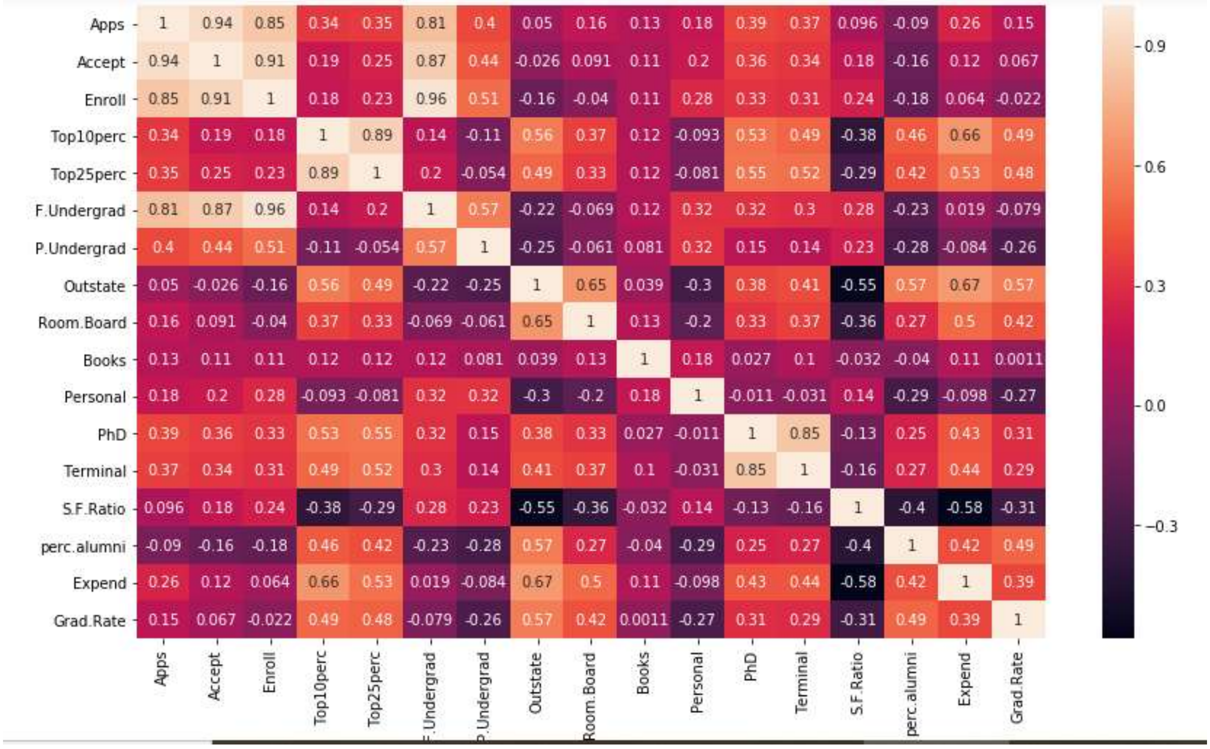
- No missing values found in the dataset
- No Duplicates found in the dataset



- Except for variable 'Top25perc', all variables have outliers present in the data.

Bivariate Analysis

Correlation between Variables



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1	0.94	0.85	0.34	0.35	0.81	0.4	0.05	0.16	0.13	0.18	0.39	0.37	0.096	-0.09	0.26	0.15
Accept	0.94	1	0.91	0.19	0.25	0.87	0.44	-0.026	0.091	0.11	0.2	0.36	0.34	0.18	-0.16	0.12	0.067
Enroll	0.85	0.91	1	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.064	-0.022
Top10perc	0.34	0.19	0.18	1	0.89	0.14	-0.11	0.56	0.37	0.12	-0.093	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1	0.2	-0.054	0.49	0.33	0.12	-0.081	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.2	1	0.57	-0.22	-0.069	0.12	0.32	0.32	0.3	0.28	-0.23	0.019	-0.079
P.Undergrad	0.4	0.44	0.51	-0.11	-0.054	0.57	1	-0.25	-0.061	0.081	0.32	0.15	0.14	0.23	-0.28	-0.084	-0.26
Outstate	0.05	-0.026	-0.16	0.56	0.49	-0.22	-0.25	1	0.65	0.039	-0.3	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.091	-0.04	0.37	0.33	-0.069	-0.061	0.65	1	0.13	-0.2	0.33	0.37	-0.36	0.27	0.5	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.081	0.039	0.13	1	0.18	0.027	0.1	-0.032	-0.04	0.11	0.0011
Personal	0.18	0.2	0.28	-0.093	-0.081	0.32	0.32	-0.3	-0.2	0.18	1	-0.011	-0.031	0.14	-0.29	-0.098	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.027	-0.011	1	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.3	0.14	0.41	0.37	0.1	-0.031	0.85	1	-0.16	0.27	0.44	0.29
S.F.Ratio	0.096	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.032	0.14	-0.13	-0.16	1	-0.4	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.4	1	0.42	0.49
Expend	0.26	0.12	0.064	0.66	0.53	0.019	-0.084	0.67	0.5	0.11	-0.098	0.43	0.44	-0.58	0.42	1	0.39
Grad.Rate	0.15	0.067	-0.022	0.49	0.48	-0.079	-0.26	0.57	0.42	0.0011	-0.27	0.31	0.29	-0.31	0.49	0.39	1

Looking at the heatmap we can say that:

- Number of applications accepted is highly correlated with number of applications received
- Number of new students enrolled is highly correlated with number of applications accepted
- Number of Full Time undergraduate students is highly correlated with number of students enrolled
- Percentage of faculties with terminal degree is highly correlated with percentage of faculties with PH.D.s'

We can also observe negative correlations between

- S.F.Ratio and Outstate S.F.Ratio
- Expend S.F.Ratio and Top10perc

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

Since the dataset have variables with different units (such as Number, Percentage and Cost),we will scale the data using zscore to standardize.

The result of **standardization** (or **Z-score normalization**) is that the features will be rescaled so that they'll have the properties of a standard normal distribution with

$$\mu=0 \text{ and } \sigma=1$$

where μ is the mean (average) and σ is the standard deviation from the mean; standard scores (also called ***z*** scores) of the samples are calculated as follows:

$$z=(x-\mu)/\sigma$$

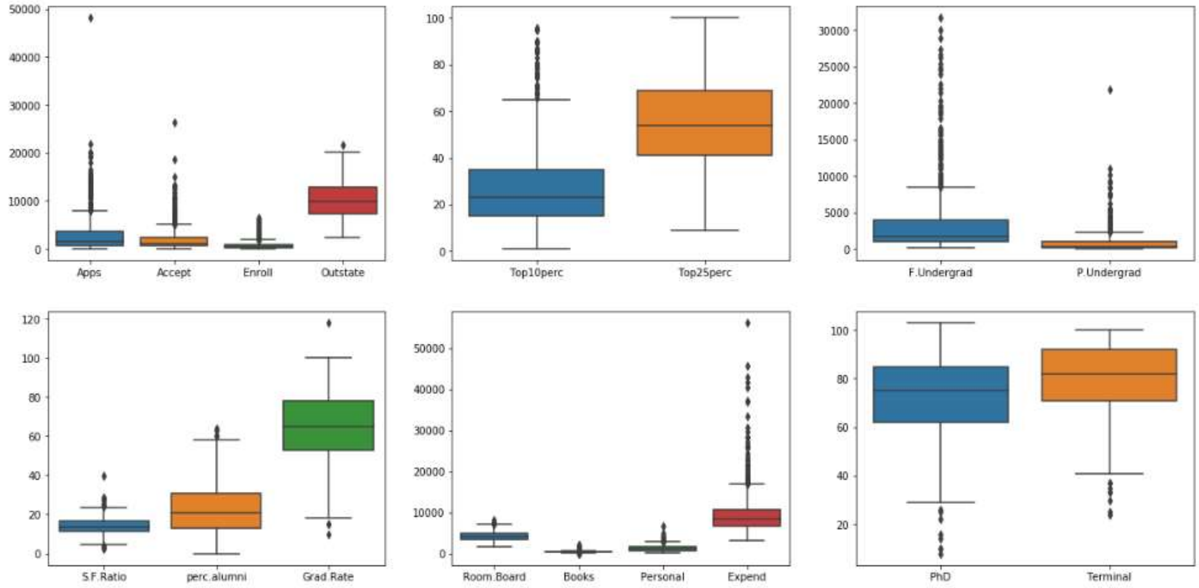
2.3 Comment on the comparison between covariance and the correlation matrix.

Covariance and Correlation matrix, both the terms measure the relationship and the dependency between two variables. "Covariance" indicates the direction of the linear relationship between variables. "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables.

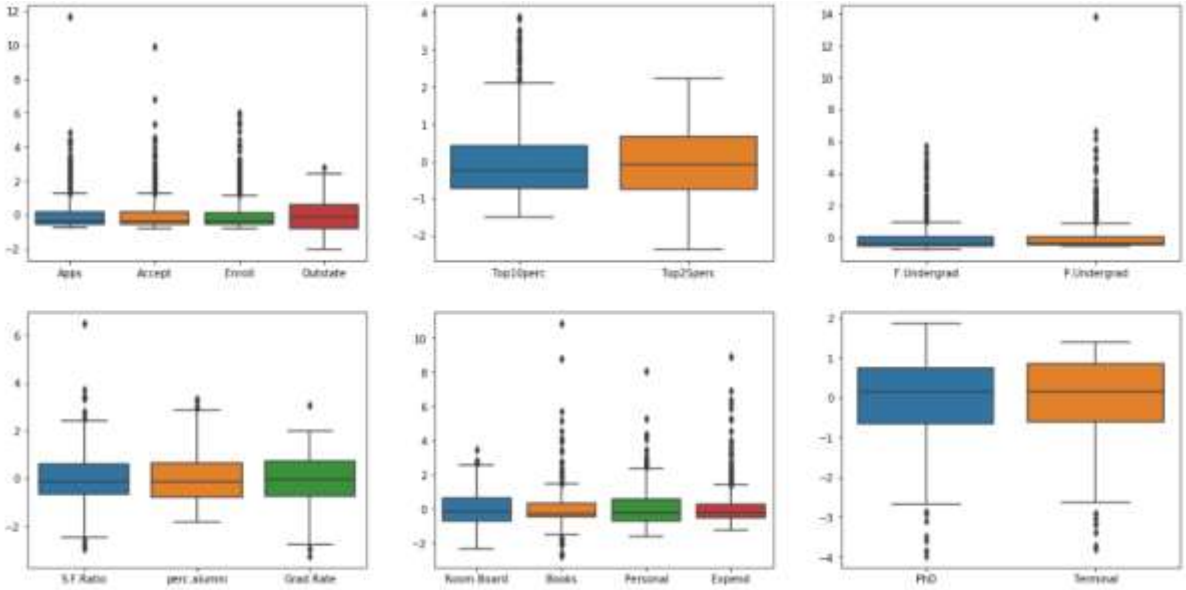
Covariance matrix should be used when the variable are on similar scales and correlation matrix when the scales of the variables differ.

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

Boxplot before Scaling:



Boxplot after Scaling:

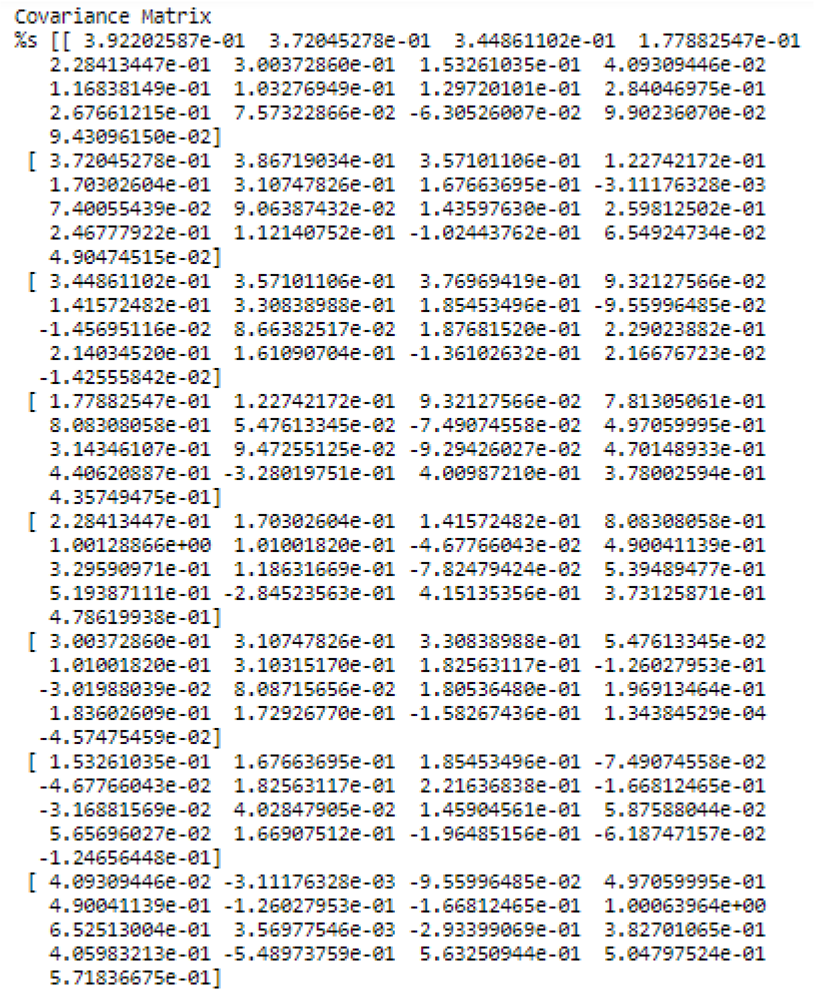


After Comparing the boxplots before and after scaling, we can see that the outliers still remain and there is no significant different on the outliers by Scaling. The outliers will have to be treated separately.

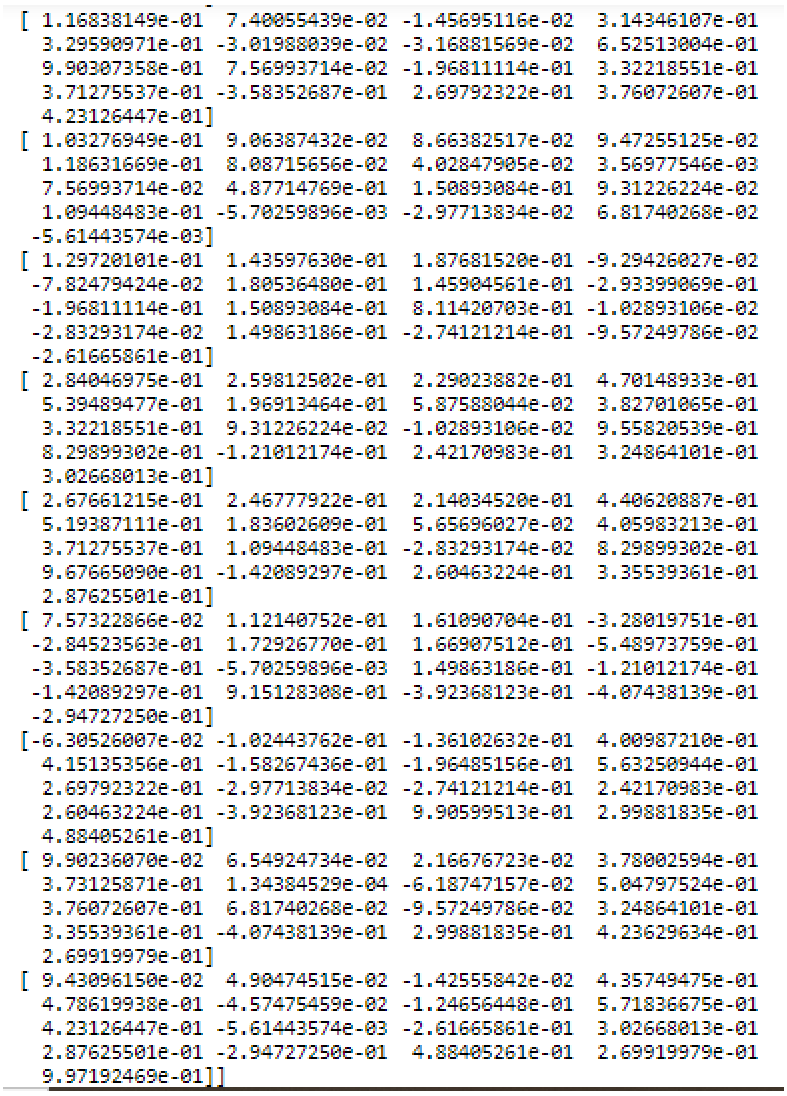
In our dataset, we have used IQR to impute the outliers detected.

2.5 Build the covariance matrix, eigenvalues and eigenvector.

Covariance Matrix

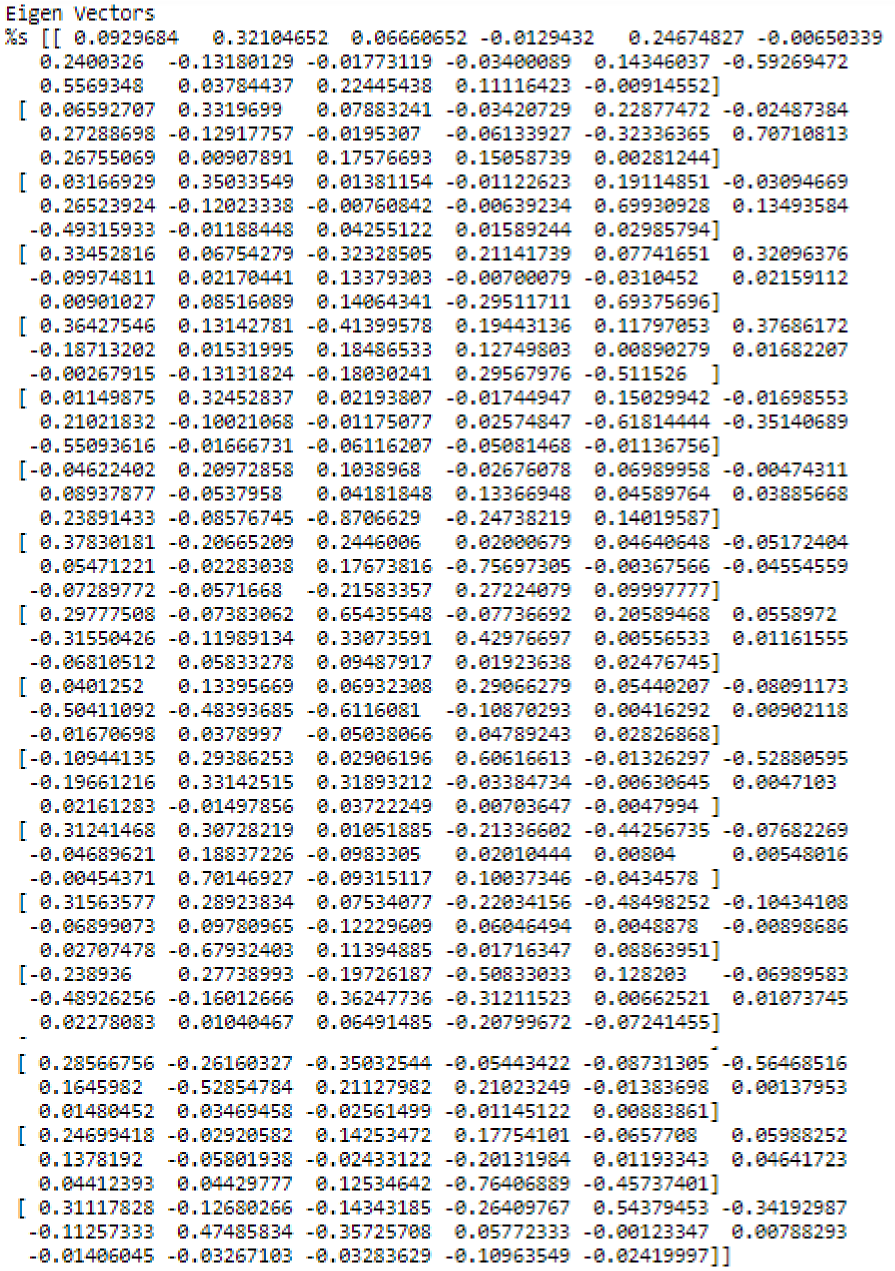


```
Covariance Matrix
%s [[ 3.92202587e-01  3.72045278e-01  3.44861102e-01  1.77882547e-01
   2.28413447e-01  3.00372860e-01  1.53261035e-01  4.09309446e-02
   1.16838149e-01  1.03276949e-01  1.29720101e-01  2.84046975e-01
   2.67661215e-01  7.57322866e-02 -6.30526007e-02  9.90236070e-02
   9.43096150e-02]
 [ 3.72045278e-01  3.86719034e-01  3.57101106e-01  1.22742172e-01
   1.70302604e-01  3.10747826e-01  1.67663695e-01 -3.11176328e-03
   7.40055439e-02  9.06387432e-02  1.43597630e-01  2.59812502e-01
   2.46777922e-01  1.12140752e-01 -1.02443762e-01  6.54924734e-02
   4.90474515e-02]
 [ 3.44861102e-01  3.57101106e-01  3.76969419e-01  9.32127566e-02
   1.41572482e-01  3.30838988e-01  1.85453496e-01 -9.55996485e-02
  -1.45695116e-02  8.66382517e-02  1.87681520e-01  2.29023882e-01
   2.14034520e-01  1.61090704e-01 -1.36102632e-01  2.16676723e-02
  -1.42555842e-02]
 [ 1.77882547e-01  1.22742172e-01  9.32127566e-02  7.81305061e-01
   8.08308058e-01  5.47613345e-02 -7.49074558e-02  4.97059995e-01
   3.14346107e-01  9.47255125e-02 -9.29426027e-02  4.70148933e-01
   4.40620887e-01 -3.28019751e-01  4.00987210e-01  3.78002594e-01
   4.35749475e-01]
 [ 2.28413447e-01  1.70302604e-01  1.41572482e-01  8.08308058e-01
   1.00128866e+00  1.01001820e-01 -4.67766043e-02  4.90041139e-01
   3.29590971e-01  1.18631669e-01 -7.82479424e-02  5.39489477e-01
   5.19387111e-01 -2.84523563e-01  4.15135356e-01  3.73125871e-01
   4.78619938e-01]
 [ 3.00372860e-01  3.10747826e-01  3.30838988e-01  5.47613345e-02
   1.01001820e-01  3.10315170e-01  1.82563117e-01 -1.26027953e-01
  -3.01988039e-02  8.08715656e-02  1.80536480e-01  1.96913464e-01
   1.83602609e-01  1.72926770e-01 -1.58267436e-01  1.34384529e-04
  -4.57475459e-02]
 [ 1.53261035e-01  1.67663695e-01  1.85453496e-01 -7.49074558e-02
  -4.67766043e-02  1.82563117e-01  2.21636838e-01 -1.66812465e-01
  -3.16881569e-02  4.02847905e-02  1.45904561e-01  5.87588044e-02
   5.65696027e-02  1.66907512e-01 -1.96485156e-01 -6.18747157e-02
  -1.24656448e-01]
 [ 4.09309446e-02 -3.11176328e-03 -9.55996485e-02  4.97059995e-01
   4.90041139e-01 -1.26027953e-01 -1.66812465e-01  1.00063964e+00
   6.52513004e-01  3.56977546e-03 -2.93399069e-01  3.82701065e-01
   4.05983213e-01 -5.48973759e-01  5.63250944e-01  5.04797524e-01
   5.71836675e-01]
```



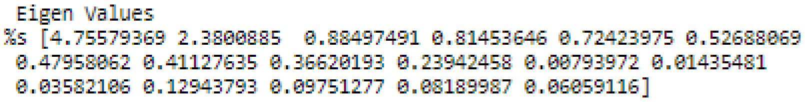
```
 [ 1.16838149e-01  7.40055439e-02 -1.45695116e-02  3.14346107e-01
   3.29590971e-01 -3.01988039e-02 -3.16881569e-02  6.52513004e-01
   9.90307358e-01  7.56993714e-02 -1.96811114e-01  3.32218551e-01
   3.71275537e-01 -3.58352687e-01  2.69792322e-01  3.76072607e-01
   4.23126447e-01]
 [ 1.03276949e-01  9.06387432e-02  8.66382517e-02  9.47255125e-02
   1.18631669e-01  8.08715656e-02  4.02847905e-02  3.56977546e-03
   7.56993714e-02  4.87714769e-01  1.50893084e-01  9.31226224e-02
   1.09448483e-01 -5.70259896e-03 -2.97713834e-02  6.81740268e-02
  -5.61443574e-03]
 [ 1.29720101e-01  1.43597630e-01  1.87681520e-01 -9.29426027e-02
  -7.82479424e-02  1.80536480e-01  1.45904561e-01 -2.93399069e-01
  -1.96811114e-01  1.50893084e-01  8.11420703e-01 -1.02893106e-02
  -2.83293174e-02  1.49863186e-01 -2.74121214e-01 -9.57249786e-02
  -2.61665861e-01]
 [ 2.84046975e-01  2.59812502e-01  2.29023882e-01  4.70148933e-01
   5.39489477e-01  1.96913464e-01  5.87588044e-02  3.82701065e-01
   3.32218551e-01  9.31226224e-02 -1.02893106e-02  9.55820539e-01
   8.29899302e-01 -1.21012174e-01  2.42170983e-01  3.24864101e-01
   3.02668013e-01]
 [ 2.67661215e-01  2.46777922e-01  2.14034520e-01  4.40620887e-01
   5.19387111e-01  1.83602609e-01  5.65696027e-02  4.05983213e-01
   3.71275537e-01  1.09448483e-01 -2.83293174e-02  8.29899302e-01
   9.67665090e-01 -1.42089297e-01  2.60463224e-01  3.35539361e-01
   2.87625501e-01]
 [ 7.57322866e-02  1.12140752e-01  1.61090704e-01 -3.28019751e-01
  -2.84523563e-01  1.72926770e-01  1.66907512e-01 -5.48973759e-01
  -3.58352687e-01 -5.70259896e-03  1.49863186e-01 -1.21012174e-01
  -1.42089297e-01  9.15128308e-01 -3.92368123e-01 -4.07438139e-01
  -2.94727250e-01]
 [-6.30526007e-02 -1.02443762e-01 -1.36102632e-01  4.00987210e-01
   4.15135356e-01 -1.58267436e-01 -1.96485156e-01  5.63250944e-01
   2.69792322e-01 -2.97713834e-02 -2.74121214e-01  2.42170983e-01
   2.60463224e-01 -3.92368123e-01  9.90599513e-01  2.99881835e-01
   4.88405261e-01]
 [ 9.90236070e-02  6.54924734e-02  2.16676723e-02  3.78002594e-01
   3.73125871e-01  1.34384529e-04 -6.18747157e-02  5.04797524e-01
   3.76072607e-01  6.81740268e-02 -9.57249786e-02  3.24864101e-01
   3.35539361e-01 -4.07438139e-01  2.99881835e-01  4.23629634e-01
   2.69919979e-01]
 [ 9.43096150e-02  4.90474515e-02 -1.42555842e-02  4.35749475e-01
   4.78619938e-01 -4.57475459e-02 -1.24656448e-01  5.71836675e-01
   4.23126447e-01 -5.61443574e-03 -2.61665861e-01  3.02668013e-01
   2.87625501e-01 -2.94727250e-01  4.88405261e-01  2.69919979e-01
   9.97192469e-01]]
```

Eigen Vectors



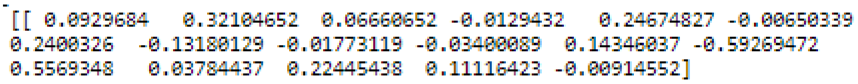
```
Eigen Vectors
%s [[ 0.0929684   0.32104652  0.06660652 -0.0129432   0.24674827 -0.00650339
   0.2400326  -0.13180129 -0.01773119 -0.03400089  0.14346037 -0.59269472
   0.5569348   0.03784437  0.22445438  0.11116423 -0.00914552]
 [ 0.06592707  0.3319699   0.07883241 -0.03420729  0.22877472 -0.02487384
   0.27288698 -0.12917757 -0.0195307  -0.06133927 -0.32336365  0.70710813
   0.26755069  0.00907891  0.17576693  0.15058739  0.00281244]
 [ 0.03166929  0.35033549  0.01381154 -0.01122623  0.19114851 -0.03094669
   0.26523924 -0.12023338 -0.00760842 -0.00639234  0.69930928  0.13493584
  -0.49315933 -0.01188448  0.04255122  0.01589244  0.02985794]
 [ 0.33452816  0.06754279 -0.32328505  0.21141739  0.07741651  0.32096376
  -0.09974811  0.02170441  0.13379303 -0.00700079 -0.0310452   0.02159112
   0.00901027  0.08516089  0.14064341 -0.29511711  0.69375696]
 [ 0.36427546  0.13142781 -0.41399578  0.19443136  0.11797053  0.37686172
  -0.18713202  0.01531995  0.18486533  0.12749803  0.00890279  0.01682207
  -0.00267915 -0.13131824 -0.18030241  0.29567976 -0.511526  ]
 [ 0.01149875  0.32452837  0.02193807 -0.01744947  0.15029942 -0.01698553
   0.21021832 -0.10021068 -0.01175077  0.02574847 -0.61814444 -0.35140689
  -0.55093616 -0.01666731 -0.06116207 -0.05081468 -0.01136756]
 [-0.04622402  0.20972858  0.1038968  -0.02676078  0.06989958 -0.00474311
   0.08937877 -0.0537958   0.04181848  0.13366948  0.04589764  0.03885668
   0.23891433 -0.08576745 -0.8706629  -0.24738219  0.14019587]
 [ 0.37830181 -0.20665209  0.2446006   0.02000679  0.04640648 -0.05172404
   0.05471221 -0.02283038  0.17673816 -0.75697305 -0.00367566 -0.04554559
  -0.07289772 -0.0571668  -0.21583357  0.27224079  0.09997777]
 [ 0.29777508 -0.07383062  0.65435548 -0.07736692  0.20589468  0.0558972
  -0.31550426 -0.11989134  0.33073591  0.42976697  0.00556533  0.01161555
  -0.06810512  0.05833278  0.09487917  0.01923638  0.02476745]
 [ 0.0401252   0.13395669  0.06932308  0.29066279  0.05440207 -0.08091173
  -0.50411092 -0.48393685 -0.6116081  -0.10870293  0.00416292  0.00902118
  -0.01670698  0.0378997  -0.05038066  0.04789243  0.02826868]
 [-0.10944135  0.29386253  0.02906196  0.60616613 -0.01326297 -0.52880595
  -0.19661216  0.33142515  0.31893212 -0.03384734 -0.00630645  0.0047103
   0.02161283 -0.01497856  0.03722249  0.00703647 -0.0047994 ]
 [ 0.31241468  0.30728219  0.01051885 -0.21336602 -0.44256735 -0.07682269
  -0.04689621  0.18837226 -0.0983305   0.02010444  0.00804     0.00548016
  -0.00454371  0.70146927 -0.09315117  0.10037346 -0.0434578 ]
 [ 0.31563577  0.28923834  0.07534077 -0.22034156 -0.48498252 -0.10434108
  -0.06899073  0.09780965 -0.12229609  0.06046494  0.0048878  -0.00898686
   0.02707478 -0.67932403  0.11394885 -0.01716347  0.08863951]
 [-0.238936    0.27738993 -0.19726187 -0.50833033  0.128203   -0.06989583
  -0.48926256 -0.16012666  0.36247736 -0.31211523  0.00662521  0.01073745
   0.02278083  0.01040467  0.06491485 -0.20799672 -0.07241455]
 [ 0.28566756 -0.26160327 -0.35032544 -0.05443422 -0.08731305 -0.56468516
   0.1645982  -0.52854784  0.21127982  0.21023249 -0.01383698  0.00137953
   0.01480452  0.03469458 -0.02561499 -0.01145122  0.00883861]
 [ 0.24699418 -0.02920582  0.14253472  0.17754101 -0.0657708   0.05988252
   0.1378192  -0.05801938 -0.02433122 -0.20131984  0.01193343  0.04641723
   0.04412393  0.04429777  0.12534642 -0.76406889 -0.45737401]
 [ 0.31117828 -0.12680266 -0.14343185 -0.26409767  0.54379453 -0.34192987
  -0.11257333  0.47485834 -0.35725708  0.05772333 -0.00123347  0.00788293
  -0.01406045 -0.03267103 -0.03283629 -0.10963549 -0.02419997]]
```

Eigen Values



```
 Eigen Values
%s [4.75579369 2.3800885  0.88497491 0.81453646 0.72423975 0.52688069
 0.47958062 0.41127635 0.36620193 0.23942458 0.00793972 0.01435481
 0.03582106 0.12943793 0.09751277 0.08189987 0.06059116]
```

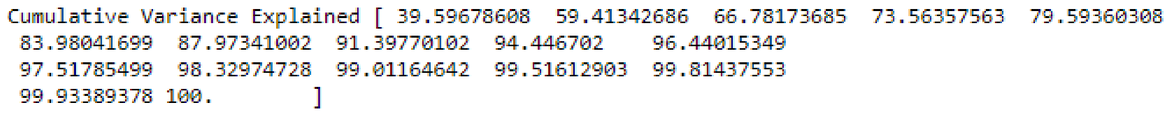
2.6 Write the explicit form of the first PC (in terms of Eigen Vectors)



```
[[ 0.0929684   0.32104652  0.06660652 -0.0129432   0.24674827 -0.00650339
0.2400326  -0.13180129 -0.01773119 -0.03400089  0.14346037 -0.59269472
0.5569348   0.03784437  0.22445438  0.11116423 -0.00914552]
```

PC1: 0.0929684*Apps + 0.32104652*Accept + 0.06660652*Enroll + -0.0129432*Top10perc + 0.24674827*Top25perc + -0.00650339*F.UnderGrad + 0.2400326*P.Undergrad + -0.13180129*Outstate + -0.01773119*Room.Board + -0.03400089*Books + 0.14346037*Personal + -0.59269472*PhD + 0.5569348*Terminal + 0.03784437*S.F.Ratio + 0.22445438*perc.alumni + 0.11116423*Expend + -0.00914552*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.



```
Cumulative Variance Explained [ 39.59678608  59.41342686  66.78173685  73.56357563  79.59360308
  83.98041699  87.97341002  91.39770102  94.446702    96.44015349
  97.51785499  98.32974728  99.01164642  99.51612903  99.81437553
  99.93389378 100.        ]
```

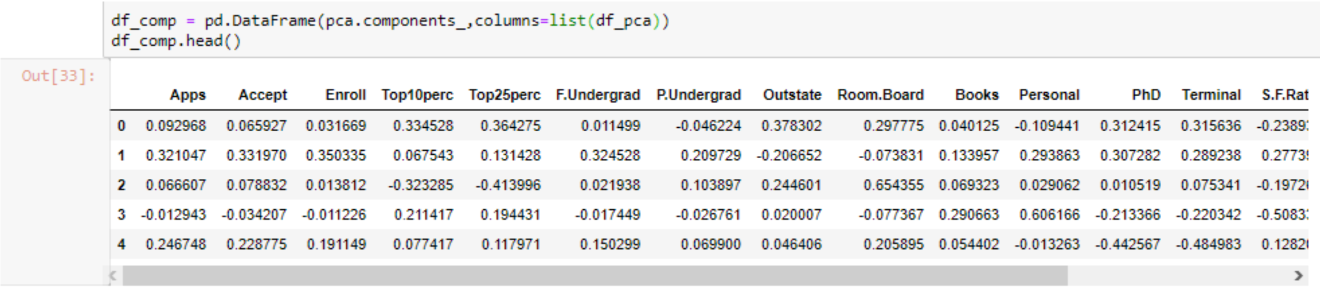
The eigenvalues tells us how much information (variance) can be attributed to each of the principal components. Cumulative Values explain how much variance is explained by the principal components together.

For instance, in this case, the first two principal components explain 59.41% of the information. Depending on the business requirement one can choose how many principal components to use in order to maximize variability.

The **eigenvectors** (principal components) determine the directions and the **eigenvalues** determine their magnitude.

In our dataset, for the dataset to contain **90% variability**, we will use **8 components**.

Principal Component scores into a dataframe:



```
df_comp = pd.DataFrame(pca.components_,columns=list(df_pca))
df_comp.head()
```

Out[33]:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
0	0.092968	0.065927	0.031669	0.334528	0.364275	0.011499	-0.046224	0.378302	0.297775	0.040125	-0.109441	0.312415	0.315636	-0.2389
1	0.321047	0.331970	0.350335	0.067543	0.131428	0.324528	0.209729	-0.206652	-0.073831	0.133957	0.293863	0.307282	0.289238	0.2773
2	0.066607	0.078832	0.013812	-0.323285	-0.413996	0.021938	0.103897	0.244601	0.654355	0.069323	0.029062	0.010519	0.075341	-0.1972
3	-0.012943	-0.034207	-0.011226	0.211417	0.194431	-0.017449	-0.026761	0.020007	-0.077367	0.290663	0.606166	-0.213366	-0.220342	-0.5083
4	0.246748	0.228775	0.191149	0.077417	0.117971	0.150299	0.069900	0.046406	0.205895	0.054402	-0.013263	-0.442567	-0.484983	0.1282

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

PCA is used to represent most accurate data in lower dimensional space. Principal Component Analysis (PCA) is a well-established mathematical technique for reducing the dimensionality of data, while keeping as much variation as possible.

In this data set we reduce the dimension from 18 to 8 variables in order to explain 90% variability in the data.