|  |
| --- |
| Project: Advanced Statistics **Business Report**  Name- Gunjar Fuley  Mobile- 9938126651 |



**Problem 1A:**

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person’s educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

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

Solution-

#One Way ANOVA (Education)

Null Hypothesis H0: Salary is not dependent on educational qualification.

Alternate Hypothesis H1: Salary is dependent on educational qualification.

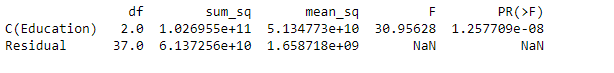
#One Way ANOVA (Occupation)

Null Hypothesis H0: Salary is not dependent on occupation.

Alternate Hypothesis H1: Salary is dependent on occupation.

1. 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.

Solution-

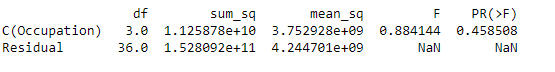


Here the PR value is much less than 0.05 therefore we reject the null hypothesis & accept the alternate hypothesis, i.e. 'Salary is dependent on educational qualification'.

At least one of the means of Salary for different Educational levels is different.

1. 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.

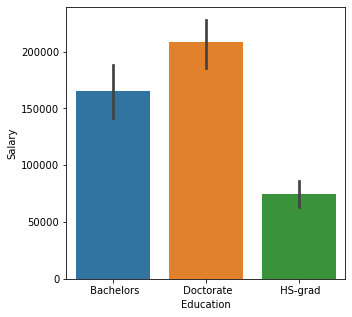
Solution-



Here the PR value is more than 0.05 therefore we accept the Null Hypothesis i.e. 'Salary is not dependent on occupation'. The means of Salary for different Occupation are equal.

1. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

Null hypothesis is rejected in (2) i.e. the ‘Education’, therefore below is the graph which shows the comparison among the means-

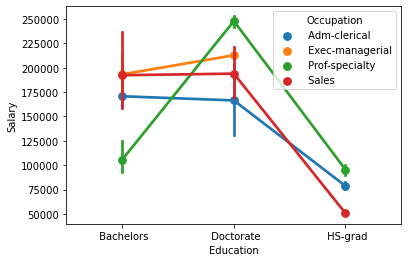


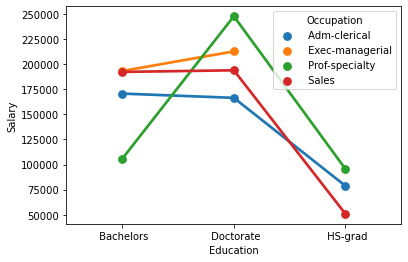
From The above graphs we can see that the there is a significant difference in the class means of ‘Doctorate Education’ & ‘HS-grad’.

**Problem 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]

Interaction effects occur when the effect of one variable depends on the value of another variable. Interaction effects are common in ANOVA.

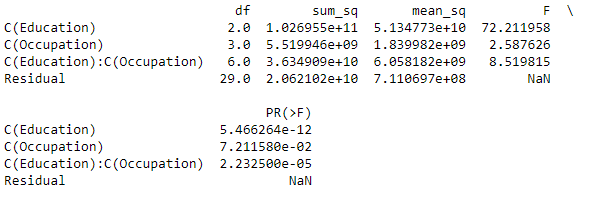




1. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education\*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

Null Hypothesis H0: Salary is not dependent on 'Education' & 'Occupation' & 'Interaction effect of Education & Occupation'.

Alternate Hypothesis H1: Salary is dependent on 'Education' & 'Occupation' & 'Interaction effect of Education & Occupation'.



Here we can see that the PR value of 'Education', 'Occupation' & 'Interaction Effect of Education & Occupation' is much low that 0.05.

Hence we reject the null hypothesis & accept the alternate hypothesis i.e.

Salary is dependent on 'Education' & 'Occupation' & 'Interaction effect of Education & Occupation'. At least one of the means are different for Educational levels, Occupation, interaction effect of Educational levels and Occupation.

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

The business implications of performing ANOVA in this case is that when we want to understand the salary is dependent of which factors, then they are combination of Education, Occupation as well as interaction between the two components. If we see individually then Education has much more significant impact on salary as compared to occupation.

**Problem 2:**

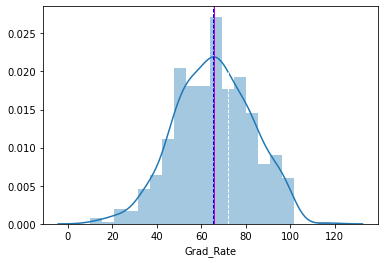
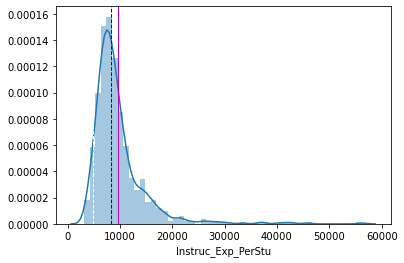
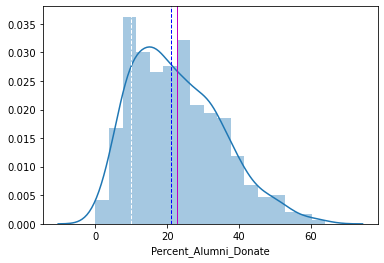
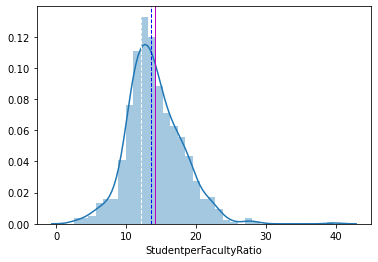
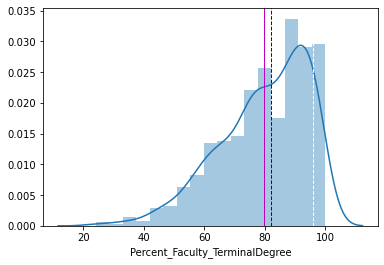
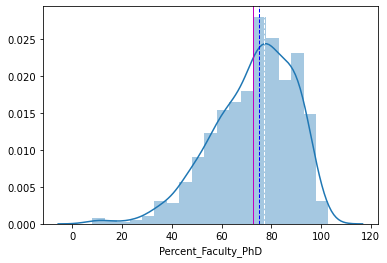
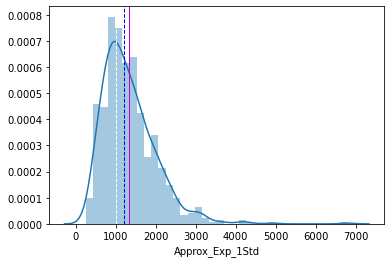
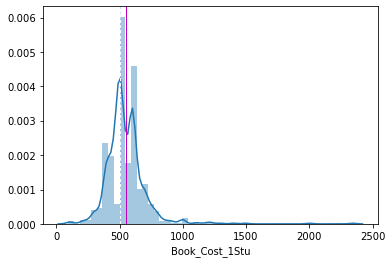
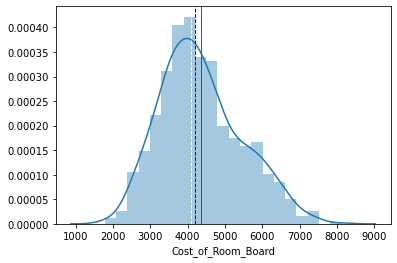
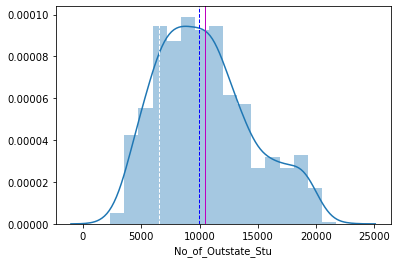
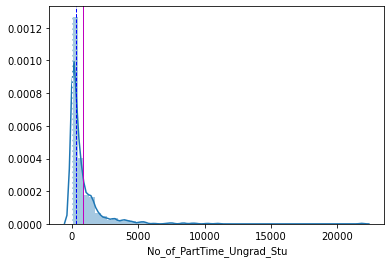
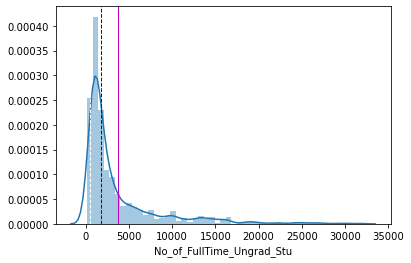
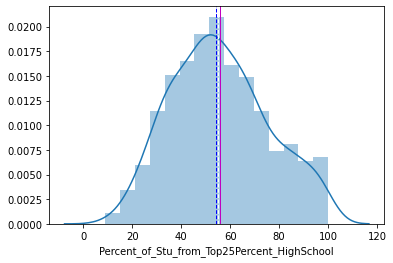
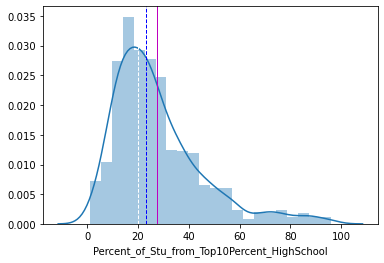
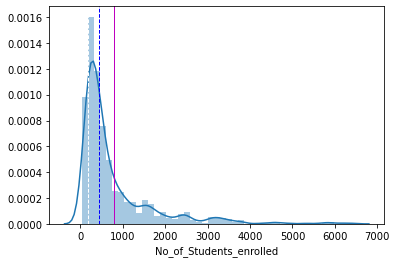
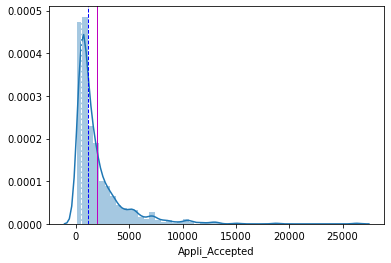
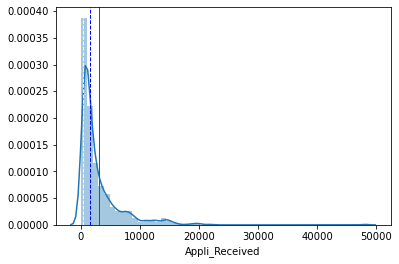
The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/29464/files/1577605/download?verifier=u1H7adFeqA94wMeycgIN6k743wS3l8LaeRBivgcN&wrap=1" \t "https://olympus.greatlearning.in/courses/29464/assignments/_blank) contains information on various colleges. You are expected 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](https://olympus.greatlearning.in/courses/29464/files/1577604/download?verifier=KWlAAbDyI9KXcOdtA2wTJyHVa8Tc6v15OxuawP6R&wrap=1" \t "https://olympus.greatlearning.in/courses/29464/assignments/_blank).

Note- I have changed the names of the variables for my better understanding-

|  |  |
| --- | --- |
| **Original Variable Name** | **New Variable Name** |
| Names | University |
| Apps | Appli\_Received |
| Accept | Appli\_Accepted |
| Enroll | No\_of\_Students\_enrolled |
| Top10perc | Percent\_of\_Stu\_from\_Top10Percent\_HighSchool |
| Top25perc | Percent\_of\_Stu\_from\_Top25Percent\_HighSchool |
| F.Undergrad | No\_of\_FullTime\_Ungrad\_Stu |
| P.Undergrad | No\_of\_PartTime\_Ungrad\_Stu |
| Outstate | No\_of\_Outstate\_Stu |
| Room.Board | Cost\_of\_Room\_Board |
| Books | Book\_Cost\_1Stu |
| Personal | Approx\_Exp\_1Std |
| PhD | Percent\_Faculty\_PhD |
| Terminal | Percent\_Faculty\_TerminalDegree |
| S.F.Ratio | StudentperFacultyRatio |
| perc.alumni | Percent\_Alumni\_Donate |
| Expend | Instruc\_Exp\_PerStu |
| Grad.Rate | Grad\_Rate |

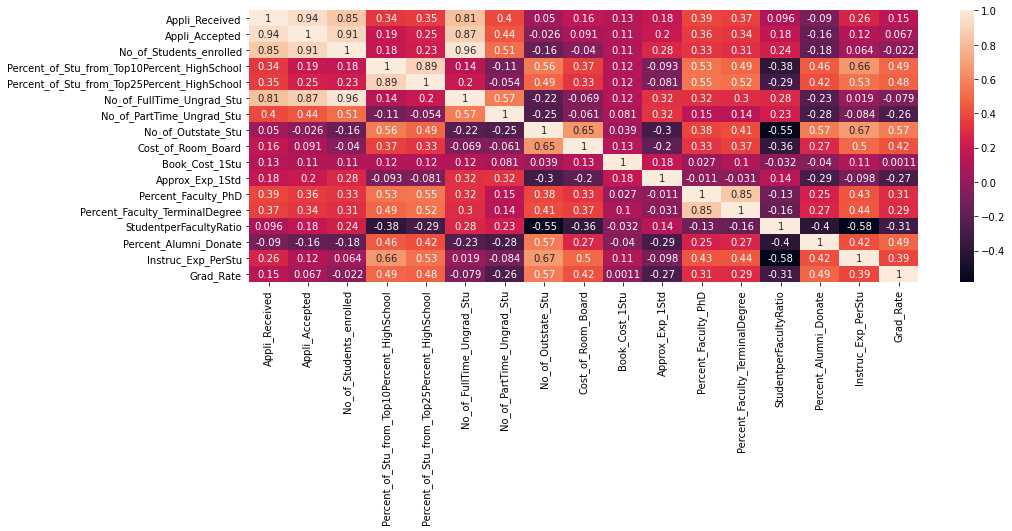
1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

Univariate Analysis



From the above distribution plots, it can be seen that Percent\_of\_Stu\_from\_Top25Percent\_HighSchool, Book\_Cost\_1Stu and Grad\_Rate looks normally distributed as mean, median and mode has approximately same values.

Multivariate Analysis



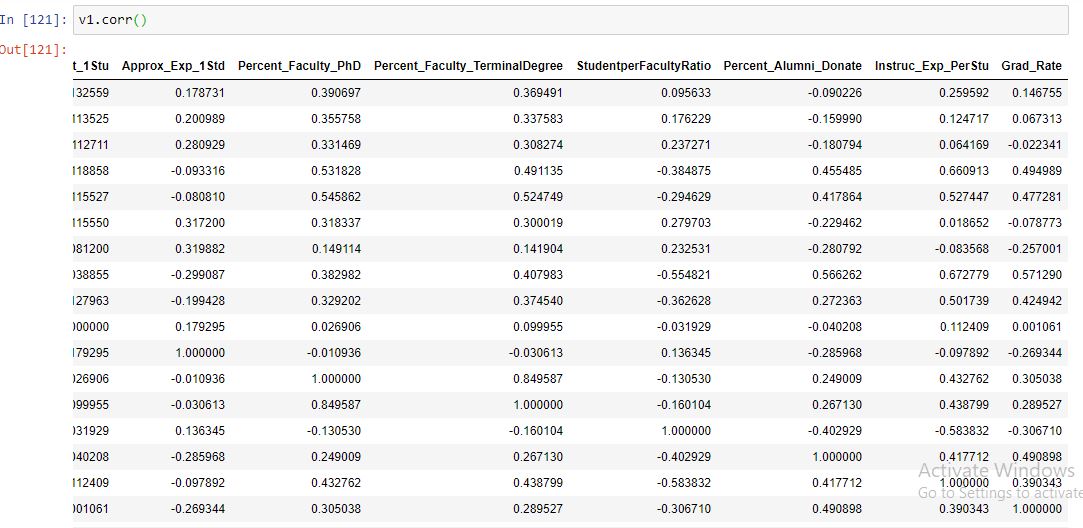
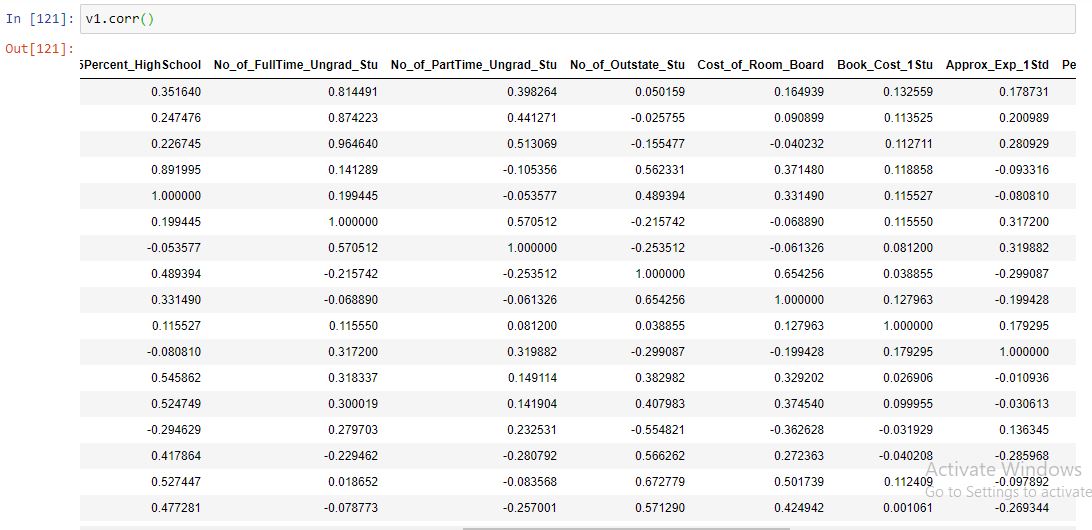
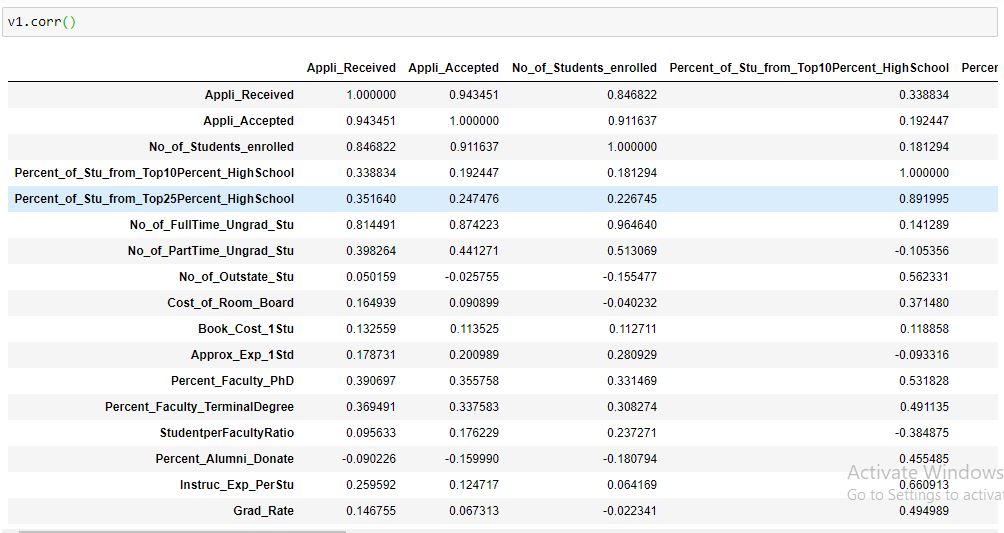
There is a high correlation between Appli\_Received and Appli\_Accepted, No\_of\_Students\_enrolled and Appli\_Accepted, No\_of\_Students\_enrolled and No\_of\_FullTime\_Ungrad\_Stu

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

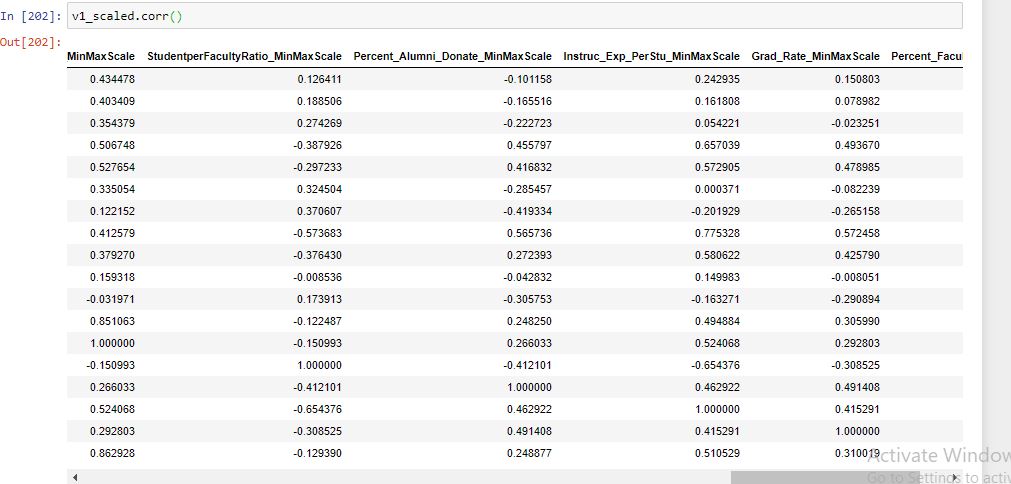
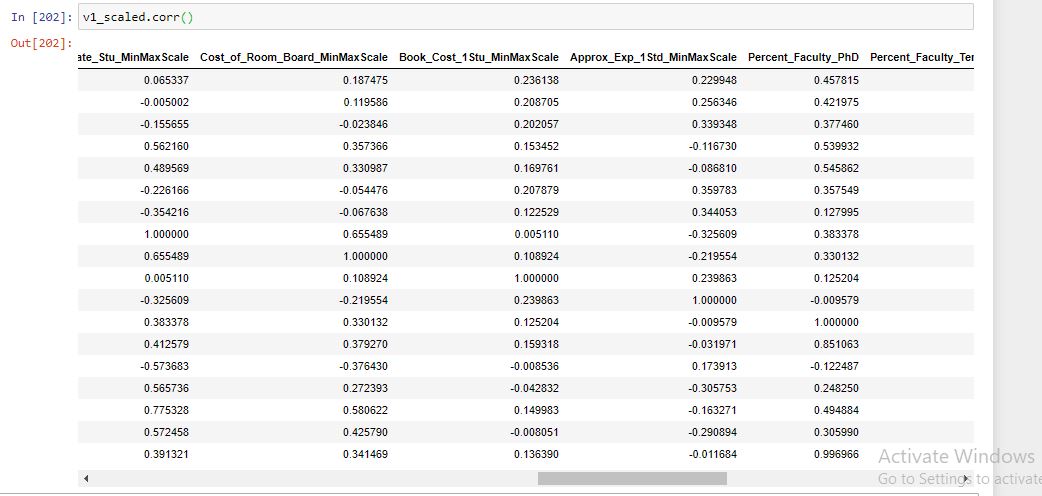
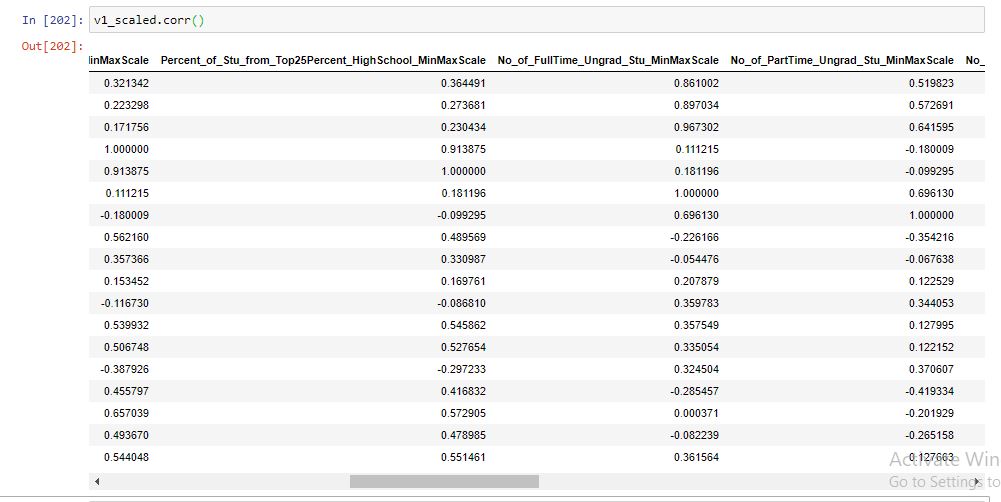
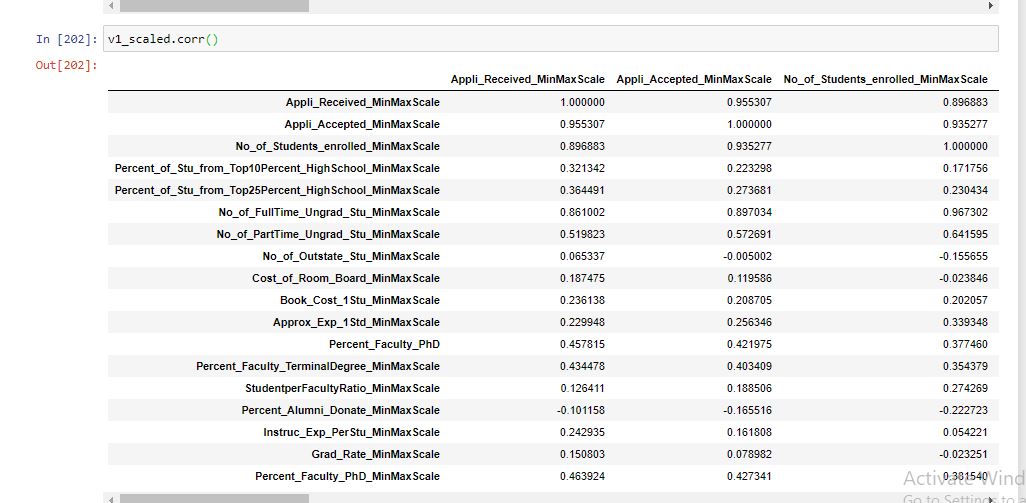
As majority of the data has outliers, and to preserve the behavior of data, the variables has been scaled using MinMaxScaler.

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

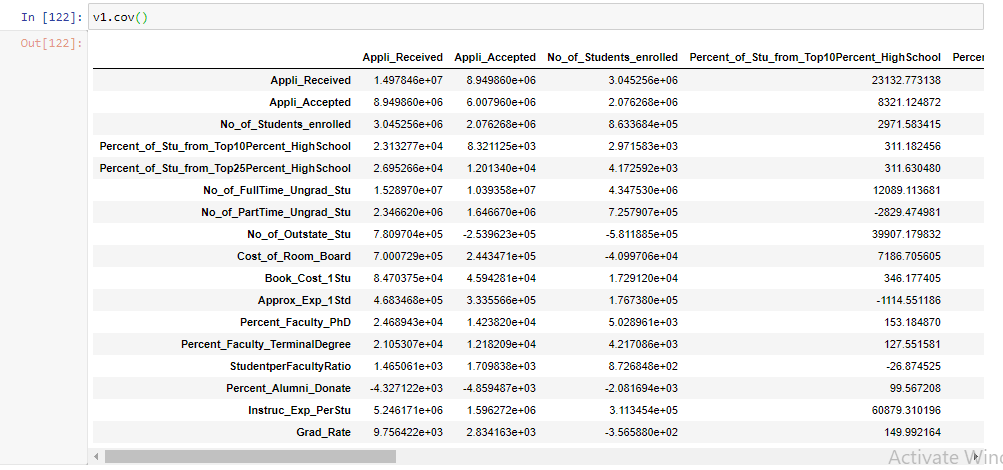
Correlation Matrix before treating outliers

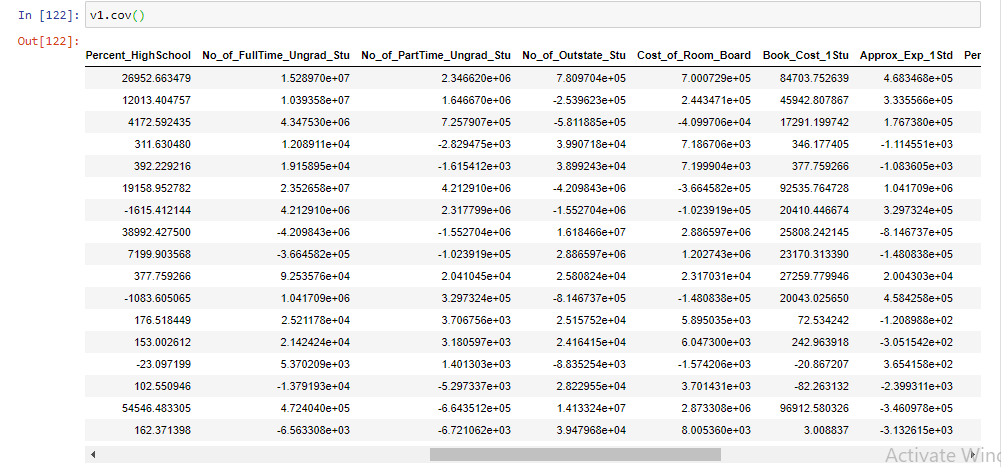


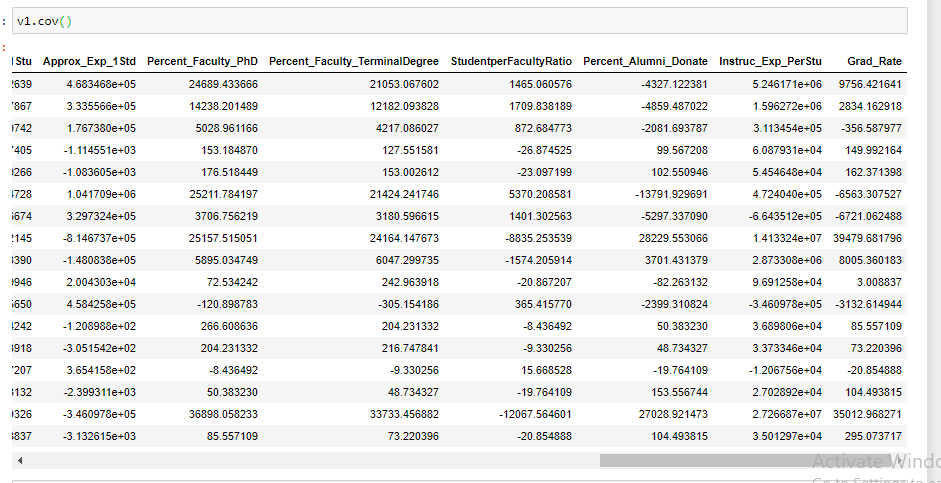
Correlation Matrix after treating outliers



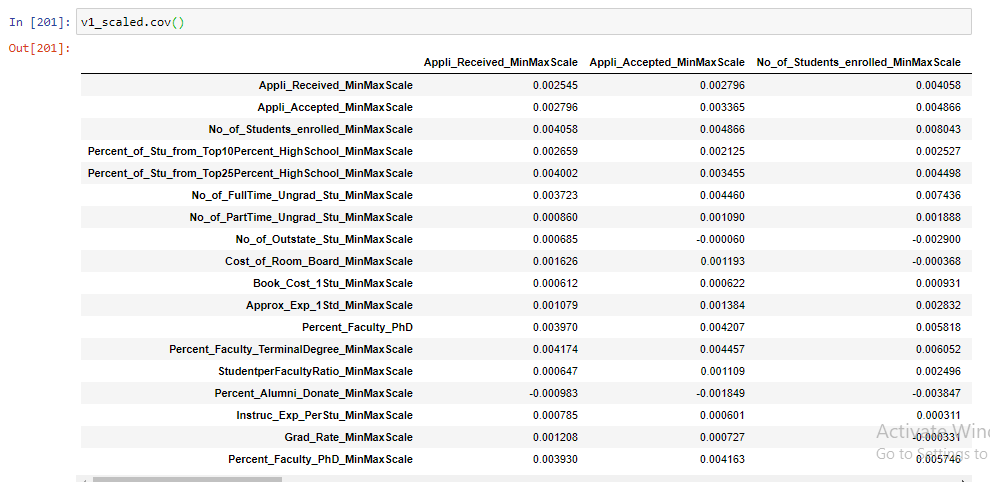
Covariance Matrix before treating outliers

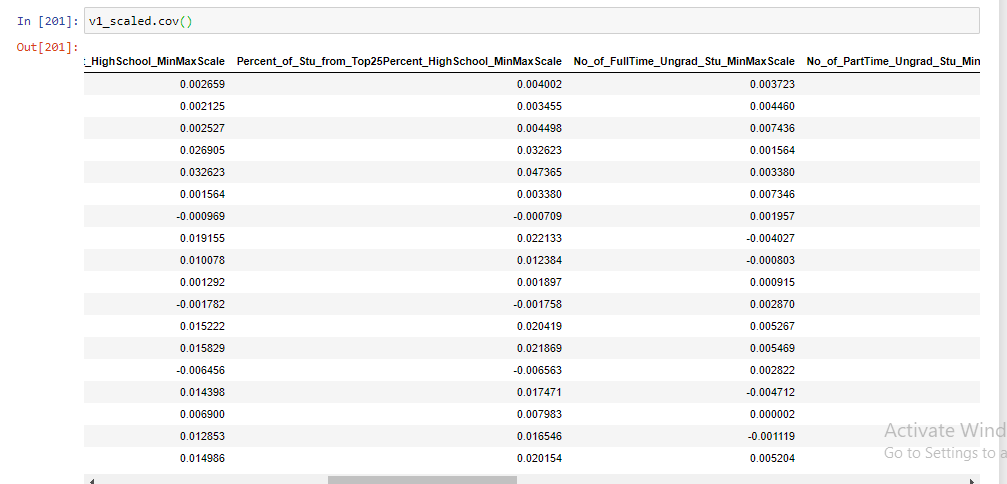


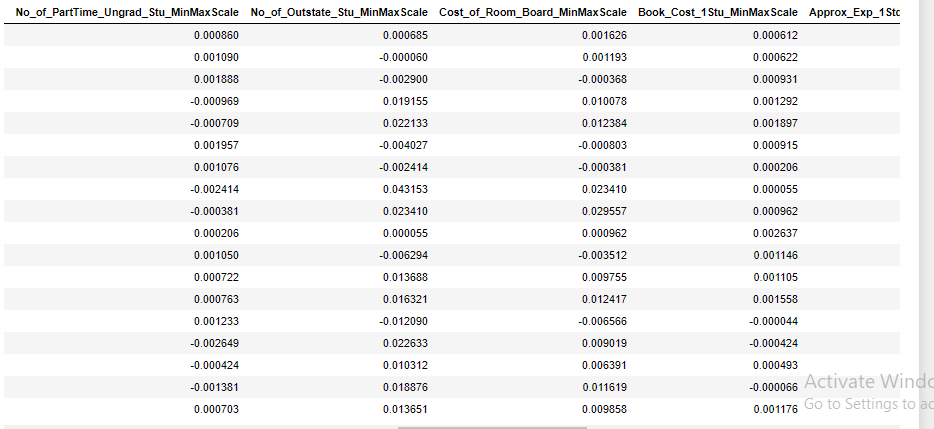


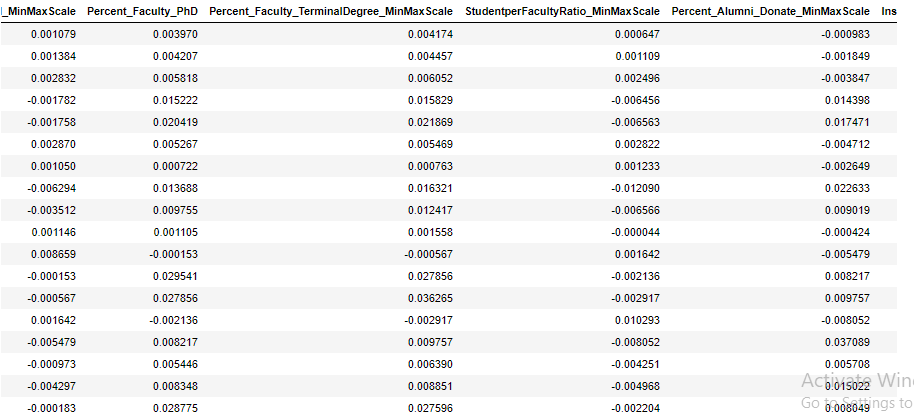


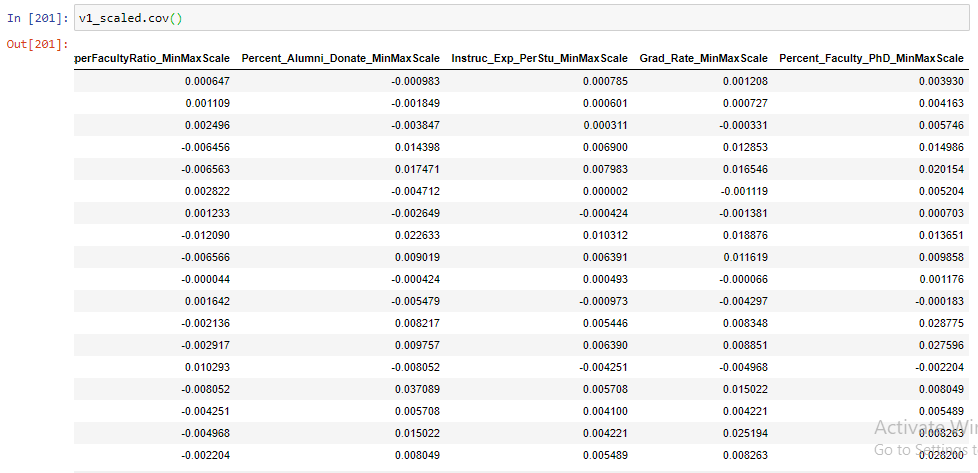
Covariance Matrix after treating outliers



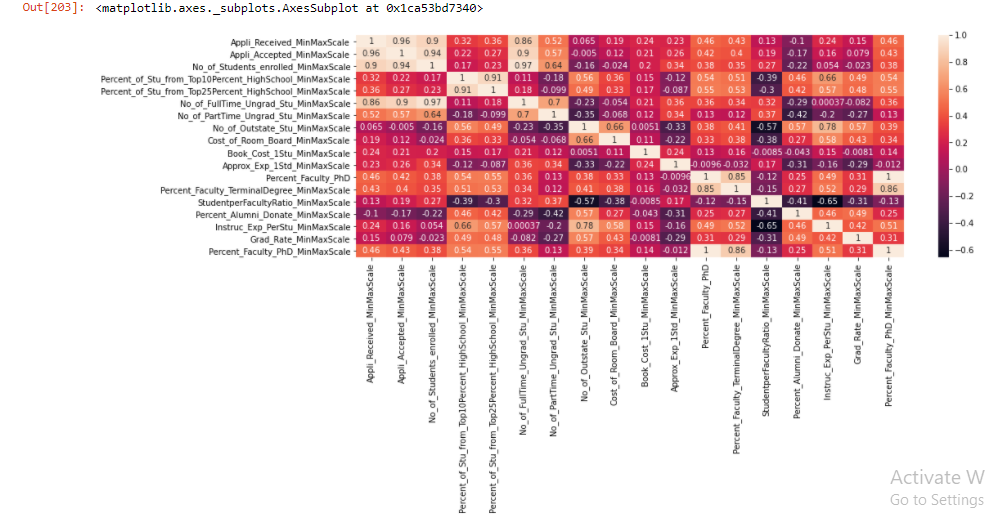








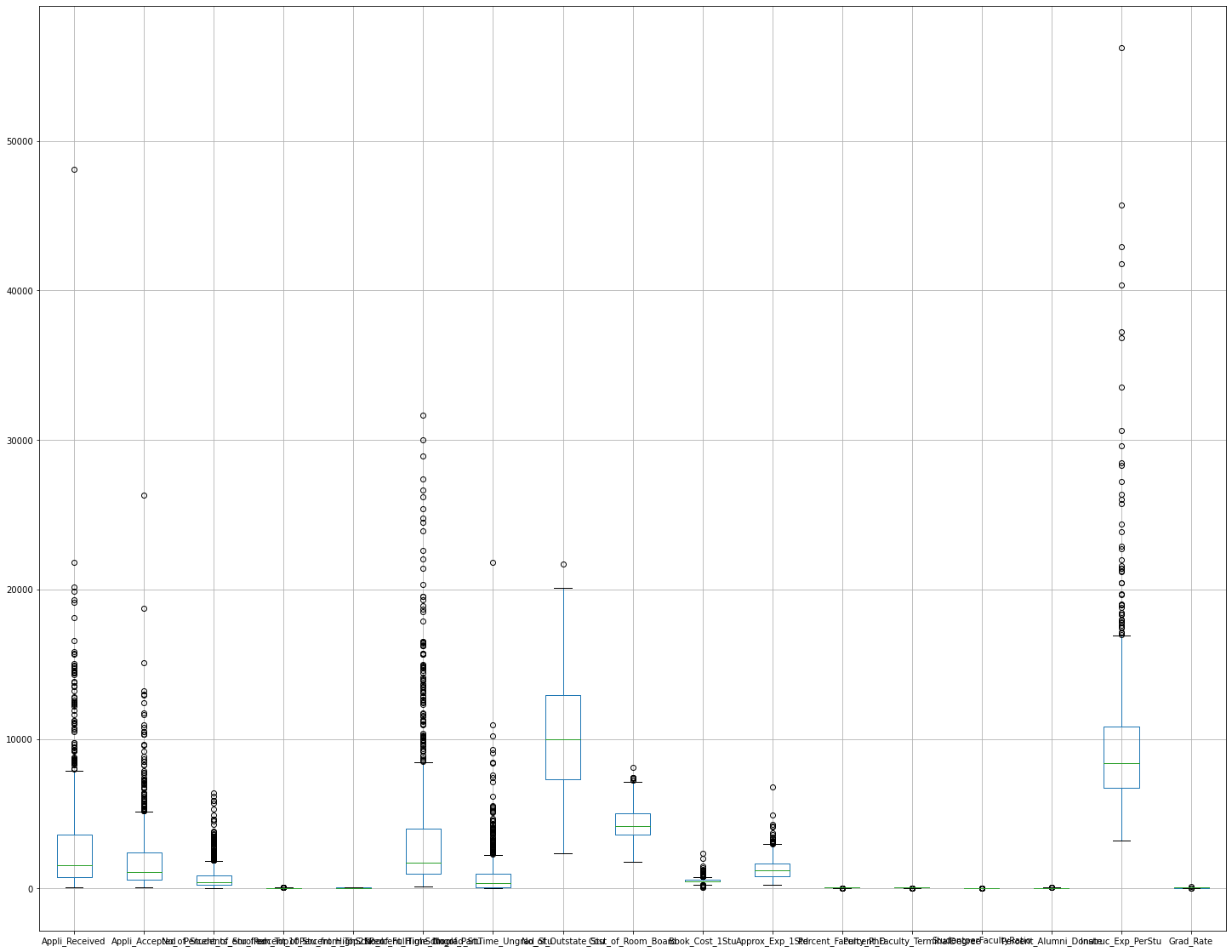
The below heat map shows the correlation after treating outliers



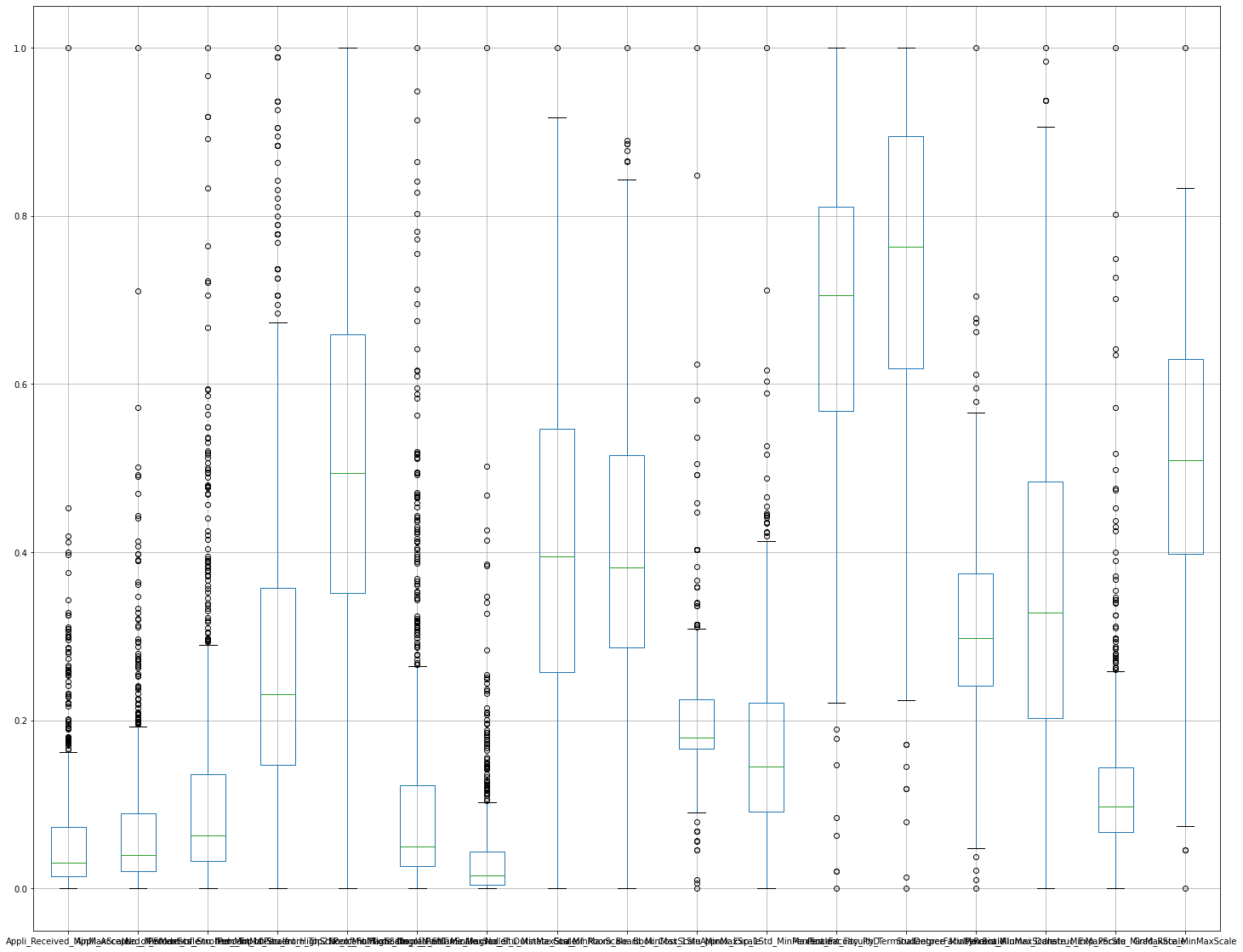
1. Check the data set for outliers before and after scaling. What insight do you derive here?

The dataset has outliers even after scaling. Scaling preserves the behavior of data.

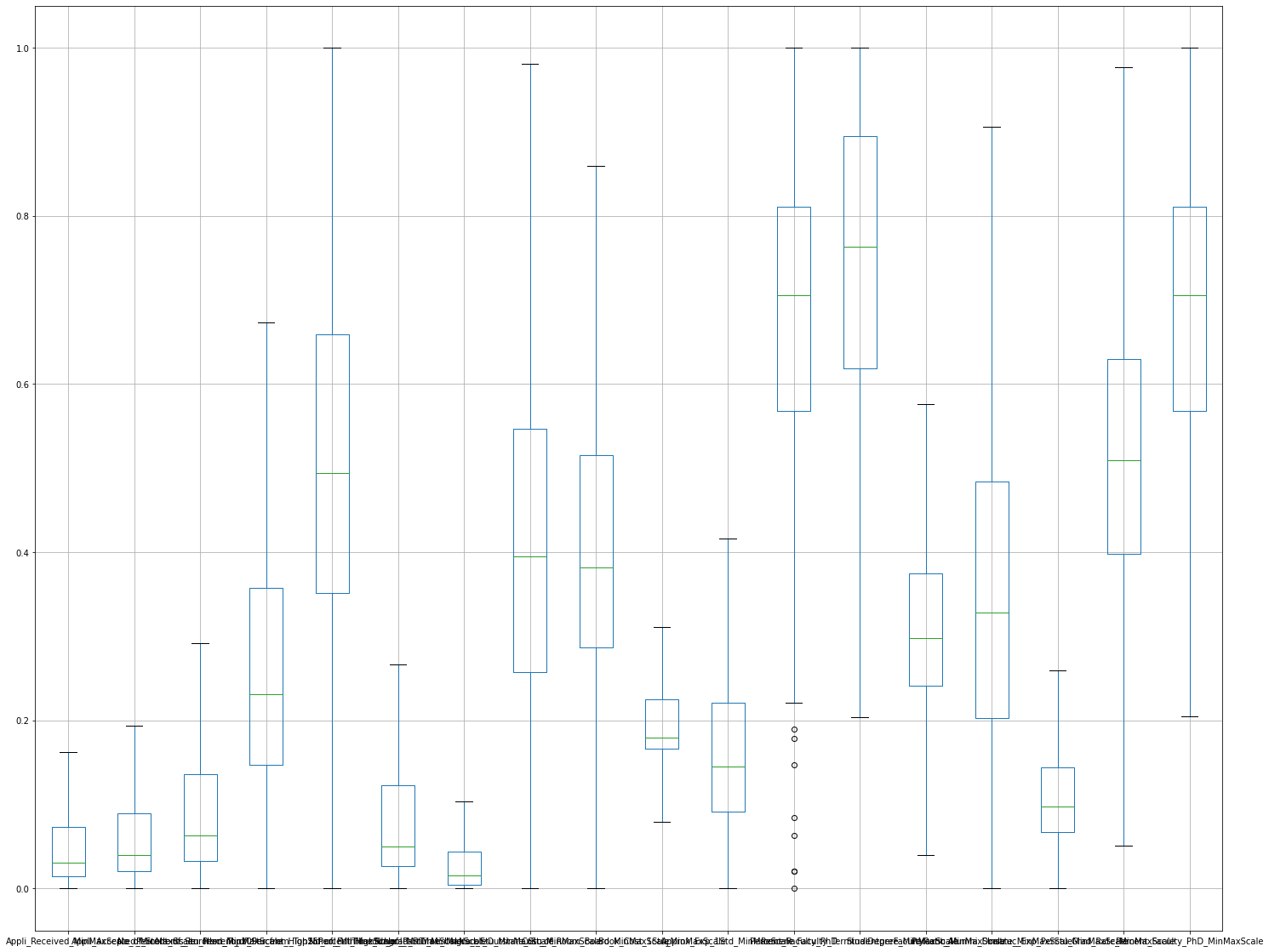
BoxPlot before scaling



After Scaling



After removal of outliers



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

array([[-0.24033092, -0.57352671, -0.24711518, ..., -0.1338771 ,

1.07467244, -0.03214799],

[-0.17881673, 0.56634086, 0.27953488, ..., -0.05432502,

0.19596405, -0.05948769],

[ 0.08860881, -0.12604174, 0.09327346, ..., 0.15397868,

0.07720293, 0.23232512],

...,

[ 0.03425038, -0.00913396, 0.05332753, ..., -0.12813839,

0.08982101, -0.12985065],

[ 0.04495219, 0.02083452, -0.0614986 , ..., -0.05277523,

0.1533457 , 0.07629538],

[ 0.02162284, -0.08239238, 0.01215097, ..., -0.00965446,

-0.04446377, -0.02428578]])

array([[ 4.49518901e-02, 3.97098744e-02, 3.95651260e-02,

3.30235869e-01, 4.31733008e-01, 2.53699041e-02,

-1.15556082e-02, 3.82200047e-01, 2.46410511e-01,

1.63405725e-02, -4.53371195e-02, 3.23792165e-01,

3.54718791e-01, -1.06864202e-01, 2.71780158e-01,

1.19558264e-01, 2.39269606e-01, 3.19371202e-01],

[-1.12611242e-01, -1.40877642e-01, -2.49133432e-01,

1.71961454e-02, -4.74204826e-02, -2.53066064e-01,

-7.93530063e-02, 3.94729693e-01, 1.83874467e-01,

-3.71406533e-02, -1.54649465e-01, -3.55749726e-01,

-3.50348804e-01, -1.89795684e-01, 3.98664502e-01,

4.98480321e-02, 2.25025995e-01, -3.46733746e-01],

[ 1.73114777e-02, 8.74517048e-03, 5.91172085e-02,

4.03002148e-01, 6.11586104e-01, 4.81641137e-02,

-1.27701217e-02, -2.97444027e-01, -4.62859225e-01,

1.76642516e-02, 6.40241920e-02, -1.79831758e-01,

-2.52794427e-01, 4.16609916e-02, 1.26277118e-01,

-4.45757082e-02, 3.93843864e-02, -1.80585018e-01],

[-1.01139599e-01, -1.07185612e-01, -1.33206790e-01,

-1.66150723e-01, -2.48646960e-01, -1.26888157e-01,

-4.79175786e-02, -1.59641062e-01, -4.95630623e-01,

-5.61008926e-02, -7.72328972e-02, 2.09393933e-01,

1.96356530e-01, 3.54893202e-02, 6.67519029e-01,

-5.76762137e-02, -4.51542535e-02, 1.94562040e-01],

[ 1.39717568e-01, 1.66271214e-01, 2.47758191e-01,

-1.63354196e-01, -1.33614200e-01, 2.25187936e-01,

5.08272077e-02, -1.73326575e-01, 4.20419180e-02,

-1.47786615e-02, -3.75556417e-02, -2.42485528e-02,

-7.51048686e-02, 2.63450401e-01, 1.05477780e-01,

-1.16437350e-01, 8.11920519e-01, -2.54351935e-02],

[ 1.91079512e-01, 2.28160157e-01, 3.71386874e-01,

-5.19992699e-02, -4.19066488e-02, 3.38908462e-01,

9.67400185e-02, -9.65673739e-03, 2.80879235e-01,

9.28969306e-02, 2.94328425e-01, -1.70593593e-01,

-1.72865750e-02, 1.39238027e-02, 5.13574746e-01,

6.23364117e-02, -3.85560123e-01, -1.57111101e-01],

[ 1.11501447e-01, 1.48946583e-01, 2.22393048e-01,

1.25891135e-03, -1.94616683e-01, 1.77382980e-01,

1.22478050e-02, 5.90366265e-01, -5.77005210e-01,

-1.23804943e-02, 2.28033827e-01, 2.09720821e-03,

-8.53440735e-02, -2.35135550e-01, -1.60584283e-01,

1.36703540e-01, 7.58901475e-02, 6.66677918e-04],

[-7.38066048e-02, -1.05246260e-01, -1.38999022e-01,

2.46775952e-02, -7.82625457e-02, -1.25377836e-01,

-2.38290720e-02, -2.98834439e-01, 1.17937032e-01,

1.29182162e-01, 6.95991387e-01, 1.12669395e-01,

-9.75245102e-02, -4.93614740e-01, 1.38141557e-02,

6.67145568e-02, 2.35232679e-01, 1.04849550e-01],

[-2.47392875e-02, -2.29882618e-02, -2.63896847e-02,

-6.61440567e-02, 4.44069007e-02, -2.04017052e-02,

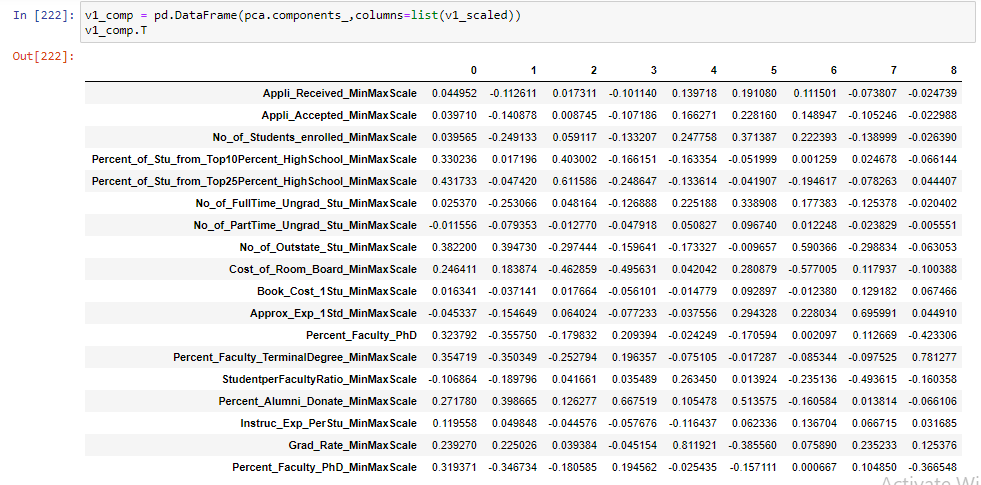
-5.55107110e-03, -6.30530221e-02, -1.00387719e-01,

6.74655879e-02, 4.49099838e-02, -4.23306203e-01,

7.81277135e-01, -1.60358024e-01, -6.61062623e-02,

3.16849182e-02, 1.25376421e-01, -3.66547584e-01]])

array([49.18, 66.64, 75.78, 82.53, 86.78, 90.08, 92.79, 94.75, 96.5 ])



1. Extract the eigenvalues and eigenvectors.[print both]

Eigen Values

%s [1.72795476e-01 6.13386713e-02 3.21238158e-02 2.37135452e-02

1.49237049e-02 1.15955040e-02 9.52482765e-03 6.87992753e-03

6.15445002e-03 5.32604385e-03 2.61861966e-03 2.24273295e-03

7.95120718e-04 5.58398685e-04 3.79314858e-04 1.91667865e-04

9.86292796e-05 7.35627733e-05]

Eigen Vectors

%s [[ 4.49518901e-02 -1.12611242e-01 1.73114777e-02 -1.01139599e-01

-1.39717568e-01 1.91079512e-01 -1.11501447e-01 7.38066048e-02

2.47392875e-02 -1.06117529e-01 1.09442456e-02 2.20729608e-02

-1.04547008e-02 -4.37437693e-01 3.90201082e-01 3.54868087e-01

6.46943550e-01 1.38409713e-02]

[ 3.97098744e-02 -1.40877642e-01 8.74517048e-03 -1.07185612e-01

-1.66271214e-01 2.28160157e-01 -1.48946583e-01 1.05246260e-01

2.29882618e-02 -1.22588370e-01 -1.04775815e-02 5.92538752e-03

-5.83300823e-02 -4.50235181e-01 3.53146419e-01 -2.29192228e-02

-7.12033486e-01 -1.26635228e-03]

[ 3.95651260e-02 -2.49133432e-01 5.91172085e-02 -1.33206790e-01

-2.47758191e-01 3.71386874e-01 -2.22393048e-01 1.38999022e-01

2.63896847e-02 -1.82422245e-01 -2.33456086e-03 -3.57018566e-02

-6.03750261e-02 6.09059646e-03 -2.77819408e-01 -6.90135555e-01

2.33722648e-01 -1.71081682e-03]

[ 3.30235869e-01 1.71961454e-02 4.03002148e-01 -1.66150723e-01

1.63354196e-01 -5.19992699e-02 -1.25891135e-03 -2.46775952e-02

6.61440567e-02 -7.08085144e-02 7.81757312e-01 -1.74323010e-01

-1.27699014e-01 4.59734425e-02 2.61972474e-02 1.07047448e-02

-1.78281844e-02 1.30945569e-02]

[ 4.31733008e-01 -4.74204826e-02 6.11586104e-01 -2.48646960e-01

1.33614200e-01 -4.19066488e-02 1.94616683e-01 7.82625457e-02

-4.44069007e-02 1.06646477e-01 -5.39556929e-01 6.95810443e-02

4.61189045e-02 -1.65570618e-02 -6.57757352e-03 -9.38366846e-03

-2.90515396e-03 -5.99507635e-03]

[ 2.53699041e-02 -2.53066064e-01 4.81641137e-02 -1.26888157e-01

-2.25187936e-01 3.38908462e-01 -1.77382980e-01 1.25377836e-01

2.04017052e-02 -1.60841055e-01 -2.44275675e-02 -1.47937586e-02

1.33400156e-02 4.06020640e-01 -3.50330760e-01 6.09101876e-01

-1.34581008e-01 -1.26308139e-02]

[-1.15556082e-02 -7.93530063e-02 -1.27701217e-02 -4.79175786e-02

-5.08272077e-02 9.67400185e-02 -1.22478050e-02 2.38290720e-02

5.55107110e-03 -2.25971961e-02 -5.35821975e-02 -2.12517527e-02

-4.02480826e-02 6.57529821e-01 7.19138526e-01 -1.49208921e-01

2.47013994e-02 2.42350552e-02]

[ 3.82200047e-01 3.94729693e-01 -2.97444027e-01 -1.59641062e-01

1.73326575e-01 -9.65673739e-03 -5.90366265e-01 2.98834439e-01

6.30530221e-02 2.97328999e-01 -7.28724532e-02 3.01197780e-02

-1.21214184e-01 4.80676773e-02 -1.71798081e-02 1.17481958e-02

1.04871202e-02 6.27173563e-03]

[ 2.46410511e-01 1.83874467e-01 -4.62859225e-01 -4.95630623e-01

-4.20419180e-02 2.80879235e-01 5.77005210e-01 -1.17937032e-01

1.00387719e-01 -1.26041707e-02 2.28653870e-02 -6.73079439e-02

-2.54423278e-02 -4.64955988e-03 -4.45684267e-02 -1.04134373e-02

-4.25111008e-03 -2.70356215e-03]

[ 1.63405725e-02 -3.71406533e-02 1.76642516e-02 -5.61008926e-02

1.47786615e-02 9.28969306e-02 1.23804943e-02 -1.29182162e-01

-6.74655879e-02 7.76233044e-02 1.78999527e-01 9.55578597e-01

-1.07141021e-01 3.40553049e-02 -1.01867978e-02 -1.99245531e-02

-5.80839330e-03 -1.68207571e-02]

[-4.53371195e-02 -1.54649465e-01 6.40241920e-02 -7.72328972e-02

3.75556417e-02 2.94328425e-01 -2.28033827e-01 -6.95991387e-01

-4.49099838e-02 5.53044055e-01 -1.88122595e-02 -1.73136897e-01

1.17453737e-02 -3.23072456e-02 3.58935849e-03 8.14722241e-03

-6.00177183e-03 4.36726696e-03]

[ 3.23792165e-01 -3.55749726e-01 -1.79831758e-01 2.09393933e-01

2.42485528e-02 -1.70593593e-01 -2.09720821e-03 -1.12669395e-01

4.23306203e-01 -3.05324571e-02 -4.31440145e-02 2.00010156e-02

-5.11359495e-02 2.50435380e-03 1.47601352e-02 -4.63692150e-03

9.18480580e-03 -6.81688954e-01]

[ 3.54718791e-01 -3.50348804e-01 -2.52794427e-01 1.96356530e-01

7.51048686e-02 -1.72865750e-02 8.53440735e-02 9.75245102e-02

-7.81277135e-01 6.75089172e-02 6.90756389e-02 -6.79813263e-02

-3.32630792e-02 -1.41623972e-02 4.76243447e-03 1.49083719e-03

5.52807617e-03 -2.29867298e-02]

[-1.06864202e-01 -1.89795684e-01 4.16609916e-02 3.54893202e-02

-2.63450401e-01 1.39238027e-02 2.35135550e-01 4.93614740e-01

1.60358024e-01 6.83148482e-01 1.94729849e-01 6.89544834e-03

2.23633548e-01 -1.76810084e-02 8.54808549e-03 -1.41717565e-02

-1.16113776e-03 -6.68148685e-03]

[ 2.71780158e-01 3.98664502e-01 1.26277118e-01 6.67519029e-01

-1.05477780e-01 5.13574746e-01 1.60584283e-01 -1.38141557e-02

6.61062623e-02 5.21484604e-04 1.37288578e-03 -6.81672264e-04

-5.28609207e-03 8.31317391e-03 1.52338165e-02 8.60048060e-03

-4.62236314e-03 3.84156929e-03]

[ 1.19558264e-01 4.98480321e-02 -4.45757082e-02 -5.76762137e-02

1.16437350e-01 6.23364117e-02 -1.36703540e-01 -6.67145568e-02

-3.16849182e-02 -1.54116478e-01 9.59890805e-02 7.80280736e-02

9.42576733e-01 9.82219538e-03 4.30177890e-02 -5.23052853e-02

-1.90913161e-02 -6.13917707e-02]

[ 2.39269606e-01 2.25025995e-01 3.93843864e-02 -4.51542535e-02

-8.11920519e-01 -3.85560123e-01 -7.58901475e-02 -2.35232679e-01

-1.25376421e-01 -1.46562861e-04 1.94728573e-02 1.18974160e-02

4.67233815e-02 3.26717144e-02 1.91516705e-03 -5.80711453e-03

-4.29616387e-03 -3.41803691e-03]

[ 3.19371202e-01 -3.46733746e-01 -1.80585018e-01 1.94562040e-01

2.54351935e-02 -1.57111101e-01 -6.66677918e-04 -1.04849550e-01

3.66547584e-01 -3.58507027e-02 -4.06886260e-02 4.95851044e-02

3.54790367e-02 -5.03998133e-03 -2.06797866e-02 -2.70229971e-03

-8.90004978e-03 7.27621266e-01]]

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

From the above Eigen vectors, explicit form for first Principal Component can be written in two ways considering the Eigen Vector as it is and with the transpose of the Eigen Vector. Both of them has components with respect to all the 17 factors.

**Explicit form 1 for PC:**

PC1 = 0.04\*Appli\_Received - 0.13\*Appli\_Accepted - 0.02\*No\_of\_Students\_enrolled -0.06\*Percent\_of\_Stu\_from\_Top10Percent\_HighSchool + 0.13\*Percent\_of\_Stu\_from\_Top25Percent\_HighSchool - 0.19\*No\_of\_FullTime\_Ungrad\_Stu + 0.11\*No\_of\_PartTime\_Ungrad\_Stu + 0.08\*No\_of\_Outstate\_Stu + 0.10\*Cost\_of\_Room\_Board - 0.01\*Book\_Cost\_1Stu - 0.01\*Approx\_Exp\_1Std - 0.021\*Percent\_Faculty\_PhD + 0.65\*Percent\_Faculty\_TerminalDegree - 0.35\*StudentperFacultyRatio - 0.39\*Percent\_Alumni\_Donate - 0.43\*Instruc\_Exp\_PerStu + 0.01\*Grad\_Rate

**Explicit form 2 for PC:**

PC1 = 0.04\*Appli\_Received + 0.03\*Appli\_Accepted + 0.03\*No\_of\_Students\_enrolled + 0.35\*Percent\_of\_Stu\_from\_Top10Percent\_HighSchool + 0.49\*Percent\_of\_Stu\_from\_Top25Percent\_HighSchool + 0.01\*No\_of\_FullTime\_Ungrad\_Stu - 0.02\*No\_of\_PartTime\_Ungrad\_Stu + 0.42\*No\_of\_Outstate\_Stu + 0.27\*Cost\_of\_Room\_Board + 0.02\*Book\_Cost\_1Stu - 0.05\*Approx\_Exp\_1Std + 0.30\*Percent\_Faculty\_PhD + 0.34\*Percent\_Faculty\_TerminalDegree - 0.12\*StudentperFacultyRatio + 0.31\*Percent\_Alumni\_Donate + 0.12844929\*Instruc\_Exp\_PerStu + 0.26\*Grad\_Rate

1. 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?

Cumulative Variance Explained [ 49.18267784 66.64146894 75.78485234 82.53442521 86.78215085

90.08257254 92.79361882 94.75184867 96.50358628 98.01953511

98.76487133 99.40321913 99.62953389 99.78847059 99.89643476

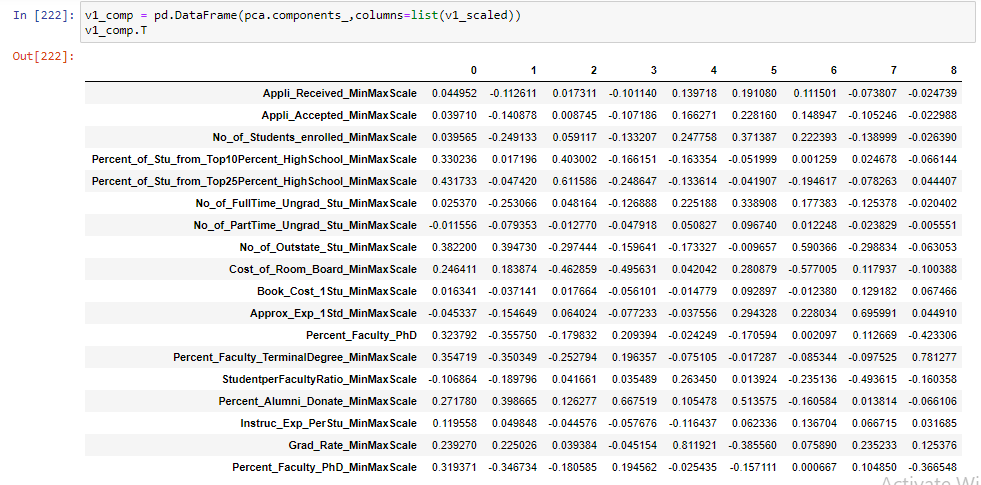
99.95098907 99.97906187 100. ]

Cumulative variance gives the per cent of features covered by the Principal Components. For example, by taking one principal components, we can cover 49.18267784% of the features and by taking two principal components, we can cover 66.64146894% of the features.

From the above figure, it can be understood that nine principal components covers more than 95% of the features, to be exact, 96.50358628%.

Eigen Vectors gives the direction of the feature with respect to all the other features. If we consider Principal Component 1, the Eigen Vector gives the direction of Principal Component 1 in all the seventeen planes or features.

1. 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]



Here total 9 PCs are chosen based on analysis.

Hence the purpose of PCA is achieved we used it to reduce dimensions.