ADVANCED STATISTICS PROJECT

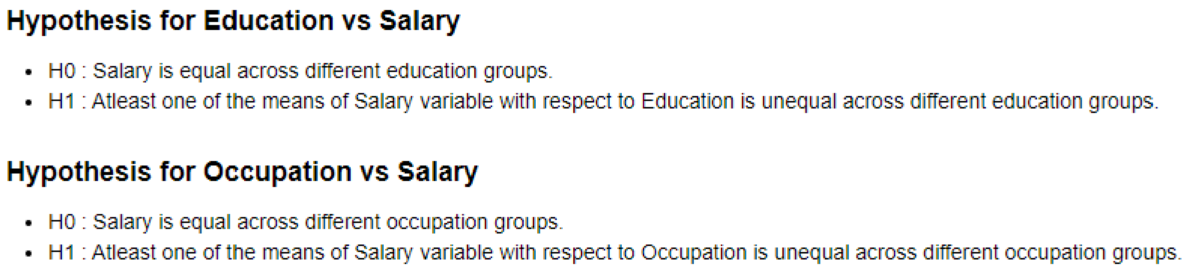
DSBA

Done By:

Hariharan Manickam

Problem 1A:

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



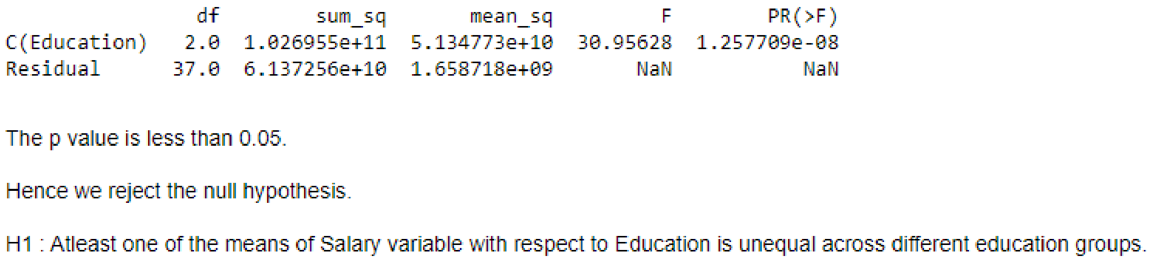
Hypothesis for Education vs Salary

- H0 : Salary is equal across different education groups.
- H1 : Atleast one of the means of Salary variable with respect to Education is unequal across different education groups.

Hypothesis for Occupation vs Salary

- H0 : Salary is equal across different occupation groups.
- H1 : Atleast one of the means of Salary variable with respect to Occupation is unequal across different occupation groups.

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.**



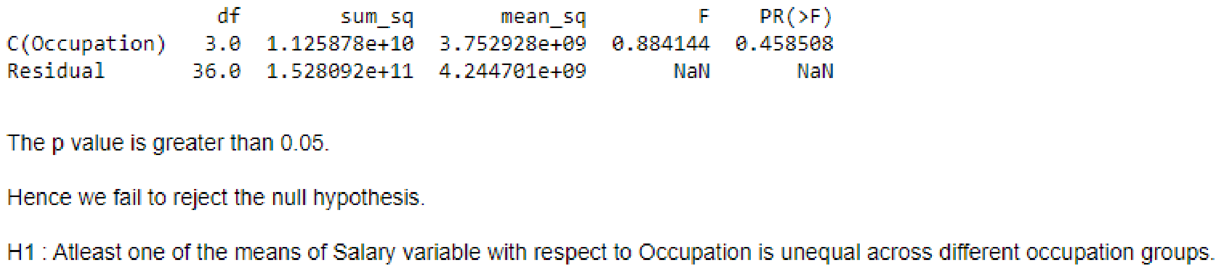
```
                 df        sum_sq       mean_sq          F        PR(>F)
C(Education)    2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual       37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

The p value is less than 0.05.

Hence we reject the null hypothesis.

H1 : Atleast one of the means of Salary variable with respect to Education is unequal across different education groups.

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.**



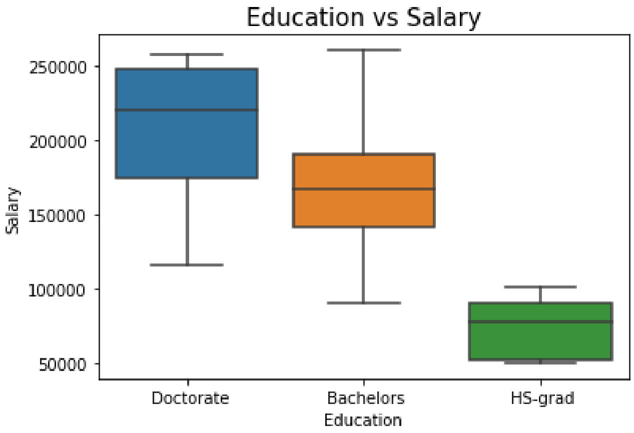
```
                  df        sum_sq       mean_sq         F    PR(>F)
C(Occupation)    3.0  1.125878e+10  3.752928e+09  0.884144  0.458508
Residual        36.0  1.528092e+11  4.244701e+09       NaN       NaN
```

The p value is greater than 0.05.

Hence we fail to reject the null hypothesis.

H1 : Atleast one of the means of Salary variable with respect to Occupation is unequal across different occupation groups.

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)

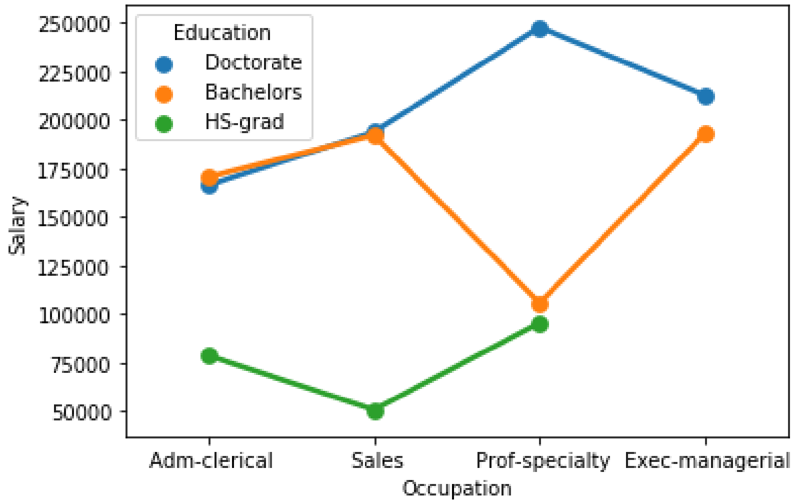


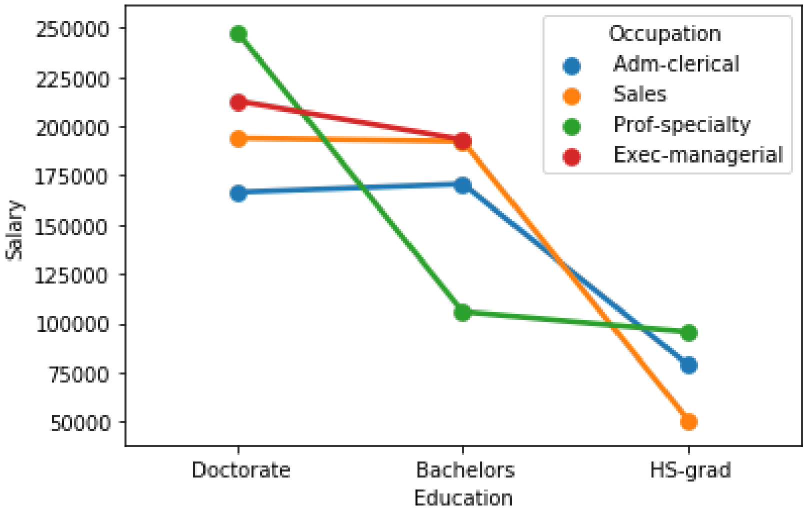
The mean salaries of all educational qualifications are varied. The mean salary of high school graduates are trhe most varied.

From the box plot, we can interpret that the salary for a person is highly dependent on their educational qualification.

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]**





From the above two interaction plots, we can observe that there is an interaction between the two variables Education and Occupation.

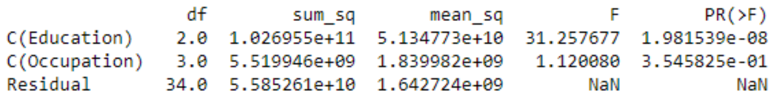
Observations

1. More the educational qualification, higher the salary
2. A person with doctorate as his/her educational qualification and works in Prof-speciality recieves the highest salary.
3. A person with HS grad as his/her educational qualification and works in Sales recieves the lowest salary.
4. The salary increase for a person working in Prof-speciality with bachelors degree and doctorate degree is huge.
5. The salary increase for a person working in sales with bachelors degree and HS grad is huge.
6. There is not much salary difference between a person with a bachelors degree and a doctorate degree working in Adm-clerical.
7. There is not much salary difference between a person with a bachelors degree and a doctorate degree working in sales.

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?

Formulate the hypothesis of ANOVA with both Education and Occupation variables with respect to the variable Salary

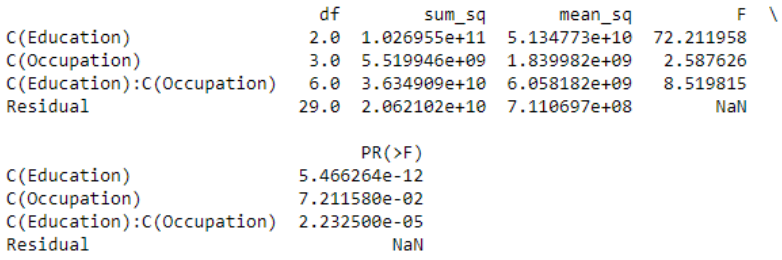
- H0 : The means of Salary with respect to the interaction of both Education and Occupation is equal.
- H1 : Atleast one of the means of Salary with respect to the interaction of both Education and Occupation category is unequal.



```
                 df        sum_sq       mean_sq          F        PR(>F)
C(Education)     2.0  1.026955e+11  5.134773e+10  31.257677  1.981539e-08
C(Occupation)    3.0  5.519946e+09  1.839982e+09   1.120080  3.545825e-01
Residual        34.0  5.585261e+10  1.642724e+09        NaN           NaN
```

The p value of Education is less than 0.05. So it is a significant factor. The p value of Occupation is greater than 0.05. So it is not a significant factor.

Next we measure the interaction effect.



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

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

We can observe that the interaction between Education and Occupation is less than 0.05. There is an interaction

Occupation as a separate individual variable is giving a different interpretation.

Hence we reject the null hypothesis.

H1 : Atleast one of the means of Salary with respect to the interaction of both Education and Occupation category is unequal.

Occupation individually does not have a significant impact on salary. Both Education & Occupation as variables in combination play a significant part in determining the salary.

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

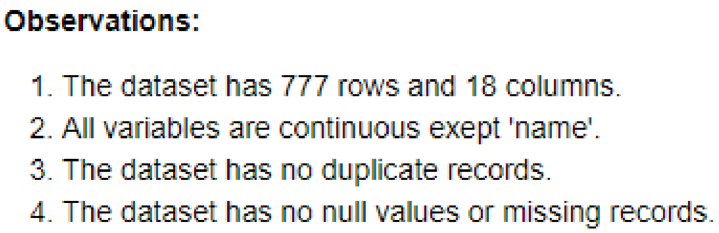
ANOVA helps us to identify the independent factors which can explain the variation obtained in the response variable.

In our case study, we observe that:

1. Education has a significant impact on Salary
2. Occupation does not have a significant impact on salary.
3. The interaction of Education & Occupation has a significant impact on salary.
4. More the educational qualification, higher the salary
5. A person with doctorate as his/her educational qualification and works in Prof-speciality recieves the highest salary.
6. A person with HS grad as his/her educational qualification and works in Sales recieves the lowest salary.
7. The salary increase for a person working in Prof-speciality with bachelors degree and doctorate degree is huge.
8. The salary increase for a person working in sales with bachelors degree and HS grad is huge.
9. There is not much salary difference between a person with a bachelors degree and a doctorate degree working in Adm-clerical.
10. There is not much salary difference between a person with a bachelors degree and a doctorate degree working in sales.

Problem 2:

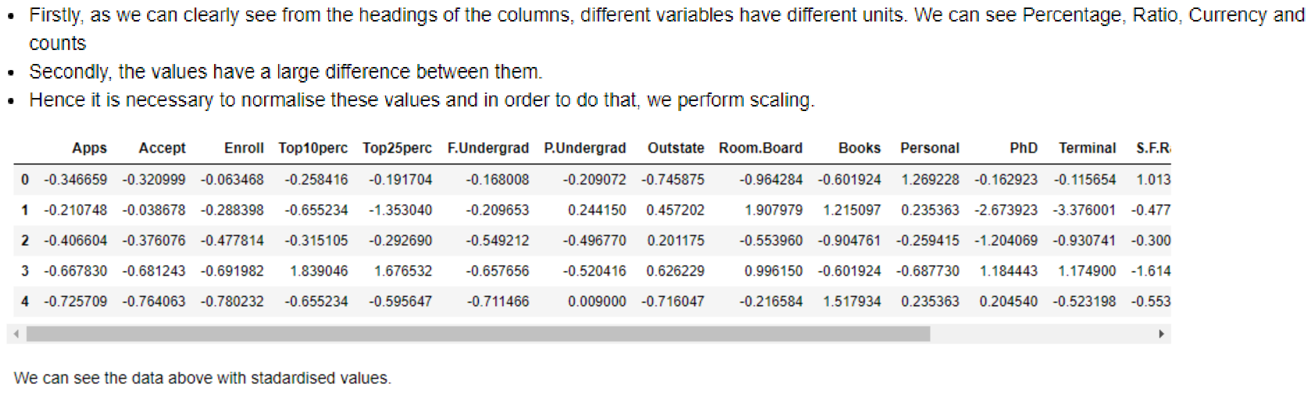
1. **Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?**



Observations:

1. The dataset has 777 rows and 18 columns.
2. All variables are continuous exept 'name'.
3. The dataset has no duplicate records.
4. The dataset has no null values or missing records.

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

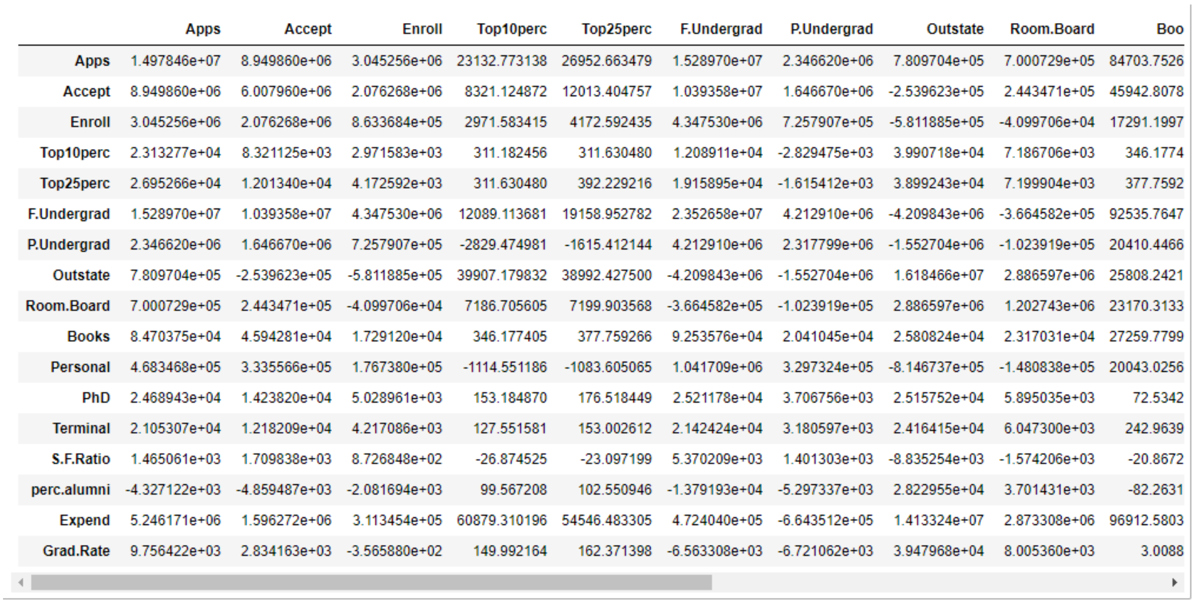


- Firstly, as we can clearly see from the headings of the columns, different variables have different units. We can see Percentage, Ratio, Currency and counts
- Secondly, the values have a large difference between them.
- Hence it is necessary to normalise these values and in order to do that, we perform scaling.

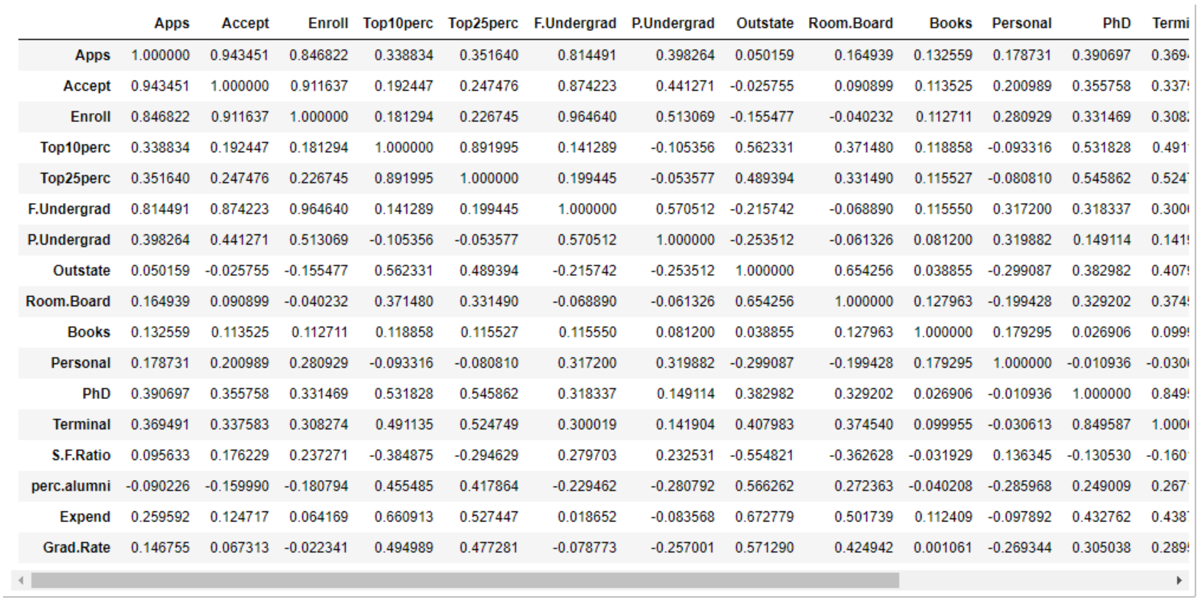
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R
0	-0.346659	-0.320999	-0.063468	-0.258416	-0.191704	-0.168008	-0.209072	-0.745875	-0.964284	-0.601924	1.269228	-0.162923	-0.115654	1.013
1	-0.210748	-0.038678	-0.288398	-0.655234	-1.353040	-0.209653	0.244150	0.457202	1.907979	1.215097	0.235363	-2.673923	-3.376001	-0.477
2	-0.406604	-0.376076	-0.477814	-0.315105	-0.292690	-0.549212	-0.496770	0.201175	-0.553960	-0.904761	-0.259415	-1.204069	-0.930741	-0.300
3	-0.667830	-0.681243	-0.691982	1.839046	1.676532	-0.657656	-0.520416	0.626229	0.996150	-0.601924	-0.687730	1.184443	1.174900	-1.614
4	-0.725709	-0.764063	-0.780232	-0.655234	-0.595647	-0.711466	0.009000	-0.716047	-0.216584	1.517934	0.235363	0.204540	-0.523198	-0.553

We can see the data above with stadardised values.

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



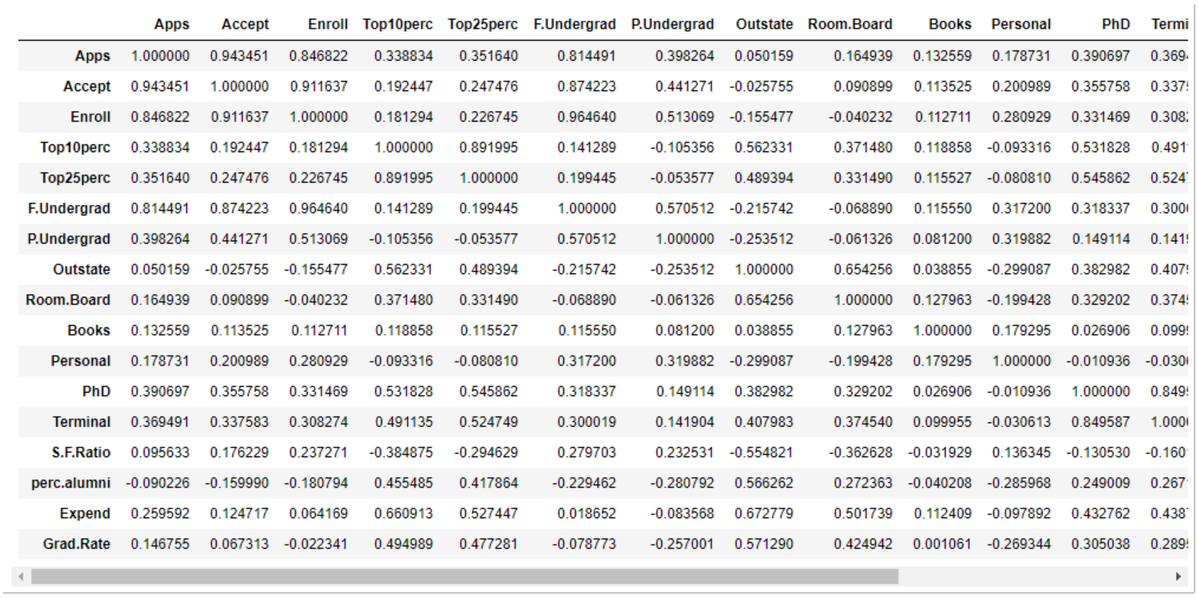
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Boo
Apps	1.497846e+07	8.949860e+06	3.045256e+06	23132.773138	26952.663479	1.528970e+07	2.346620e+06	7.809704e+05	7.000729e+05	84703.7526
Accept	8.949860e+06	6.007960e+06	2.076268e+06	8321.124872	12013.404757	1.039358e+07	1.646670e+06	-2.539623e+05	2.443471e+05	45942.8078
Enroll	3.045256e+06	2.076268e+06	8.633684e+05	2971.583415	4172.592435	4.347530e+06	7.257907e+05	-5.811885e+05	-4.099706e+04	17291.1997
Top10perc	2.313277e+04	8.321125e+03	2.971583e+03	311.182456	311.630480	1.208911e+04	-2.829475e+03	3.990718e+04	7.186706e+03	346.1774
Top25perc	2.695266e+04	1.201340e+04	4.172592e+03	311.630480	392.229216	1.915895e+04	-1.615412e+03	3.899243e+04	7.199904e+03	377.7592
F.Undergrad	1.528970e+07	1.039358e+07	4.347530e+06	12089.113681	19158.952782	2.352658e+07	4.212910e+06	-4.209843e+06	-3.664582e+05	92535.7647
P.Undergrad	2.346620e+06	1.646670e+06	7.257907e+05	-2829.474981	-1615.412144	4.212910e+06	2.317799e+06	-1.552704e+06	-1.023919e+05	20410.4466
Outstate	7.809704e+05	-2.539623e+05	-5.811885e+05	39907.179832	38992.427500	-4.209843e+06	-1.552704e+06	1.618466e+07	2.886597e+06	25808.2421
Room.Board	7.000729e+05	2.443471e+05	-4.099706e+04	7186.705605	7199.903568	-3.664582e+05	-1.023919e+05	2.886597e+06	1.202743e+06	23170.3133
Books	8.470375e+04	4.594281e+04	1.729120e+04	346.177405	377.759266	9.253576e+04	2.041045e+04	2.580824e+04	2.317031e+04	27259.7799
Personal	4.683468e+05	3.335566e+05	1.767380e+05	-1114.551186	-1083.605065	1.041709e+06	3.297324e+05	-8.146737e+05	-1.480838e+05	20043.0256
PhD	2.468943e+04	1.423820e+04	5.028961e+03	153.184870	176.518449	2.521178e+04	3.706756e+03	2.515752e+04	5.895035e+03	72.5342
Terminal	2.105307e+04	1.218209e+04	4.217086e+03	127.551581	153.002612	2.142424e+04	3.180597e+03	2.416415e+04	6.047300e+03	242.9639
S.F.Ratio	1.465061e+03	1.709838e+03	8.726848e+02	-26.874525	-23.097199	5.370209e+03	1.401303e+03	-8.835254e+03	-1.574206e+03	-20.8672
perc.alumni	-4.327122e+03	-4.859487e+03	-2.081694e+03	99.567208	102.550946	-1.379193e+04	-5.297337e+03	2.822955e+04	3.701431e+03	-82.2631
Expend	5.246171e+06	1.596272e+06	3.113454e+05	60879.310196	54546.483305	4.724040e+05	-6.643512e+05	1.413324e+07	2.873308e+06	96912.5803
Grad.Rate	9.756422e+03	2.834163e+03	-3.565880e+02	149.992164	162.371398	-6.563308e+03	-6.721062e+03	3.947968e+04	8.005360e+03	3.0088



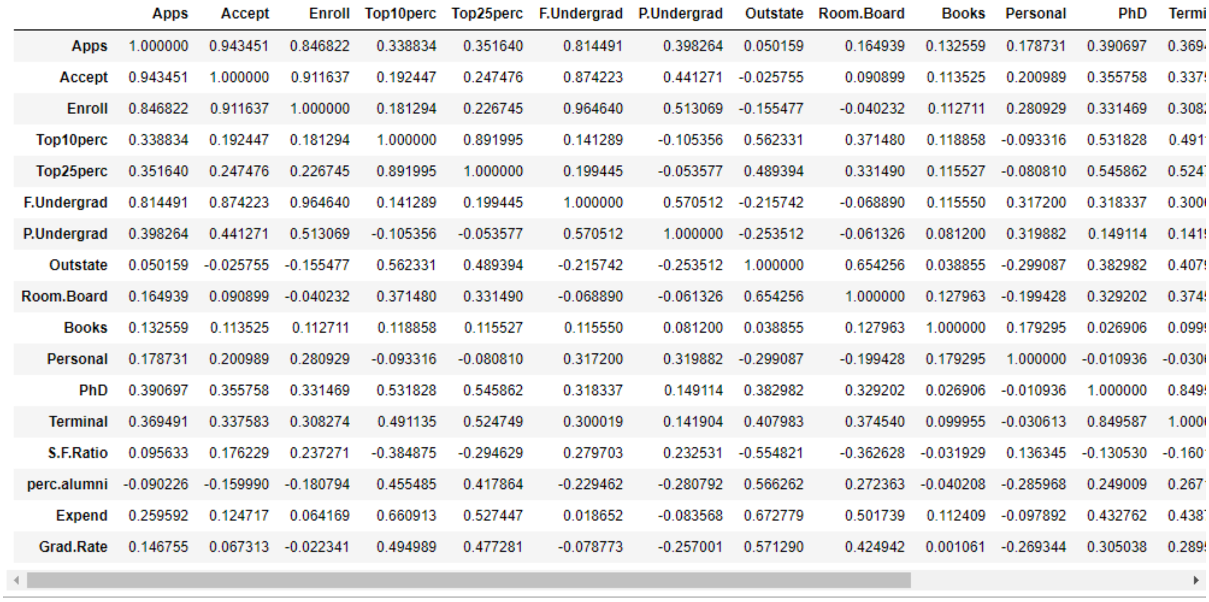
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	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

From the 2 tables above, we can say that the covariance table is not the same. This is expected as standardising changes the values.

Next we check the correlation values for both standardised and non standardised values.



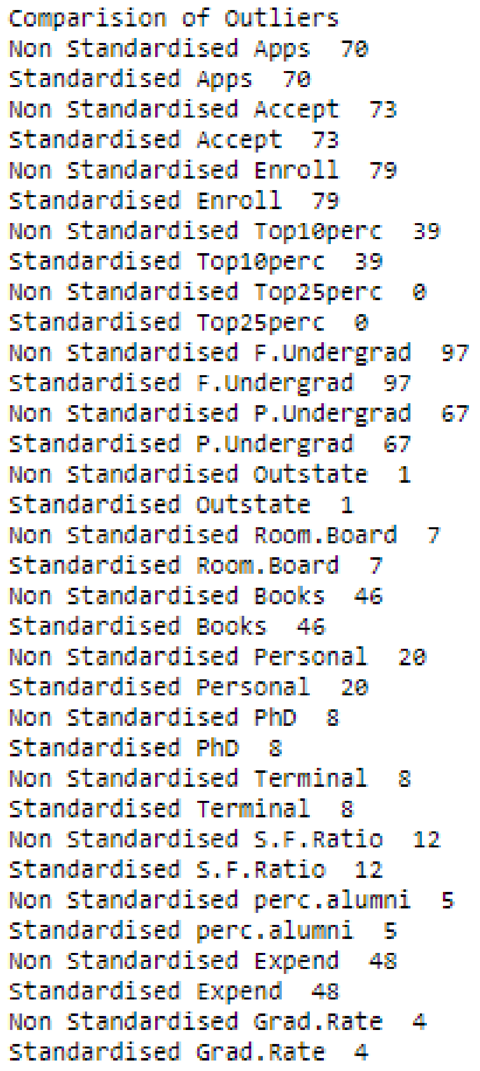
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	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



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
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

From the 2 tables above, we can say that the correlation table is the same. This is expected as standardising does not change the values.

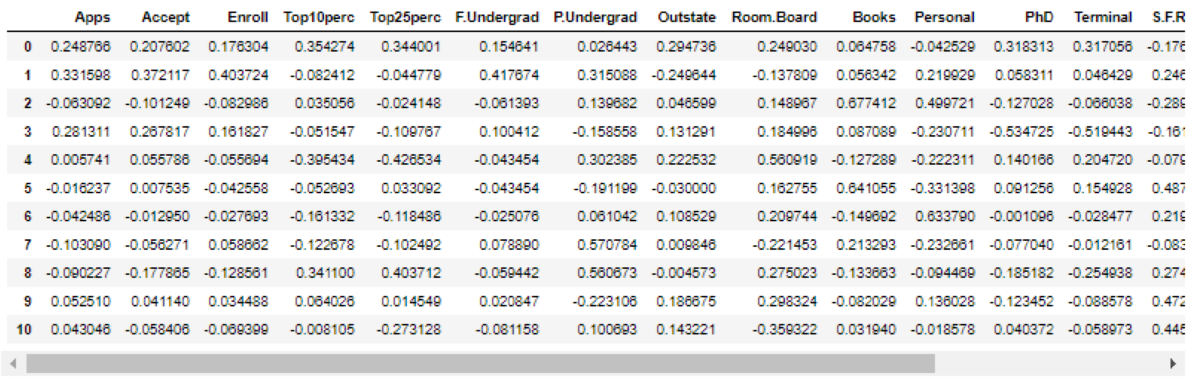
4. **Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]**



```
Comparision of Outliers
Non Standardised Apps  70
Standardised Apps  70
Non Standardised Accept  73
Standardised Accept  73
Non Standardised Enroll  79
Standardised Enroll  79
Non Standardised Top10perc  39
Standardised Top10perc  39
Non Standardised Top25perc  0
Standardised Top25perc  0
Non Standardised F.Undergrad  97
Standardised F.Undergrad  97
Non Standardised P.Undergrad  67
Standardised P.Undergrad  67
Non Standardised Outstate  1
Standardised Outstate  1
Non Standardised Room.Board  7
Standardised Room.Board  7
Non Standardised Books  46
Standardised Books  46
Non Standardised Personal  20
Standardised Personal  20
Non Standardised PhD  8
Standardised PhD  8
Non Standardised Terminal  8
Standardised Terminal  8
Non Standardised S.F.Ratio  12
Standardised S.F.Ratio  12
Non Standardised perc.alumni  5
Standardised perc.alumni  5
Non Standardised Expend  48
Standardised Expend  48
Non Standardised Grad.Rate  4
Standardised Grad.Rate  4
```

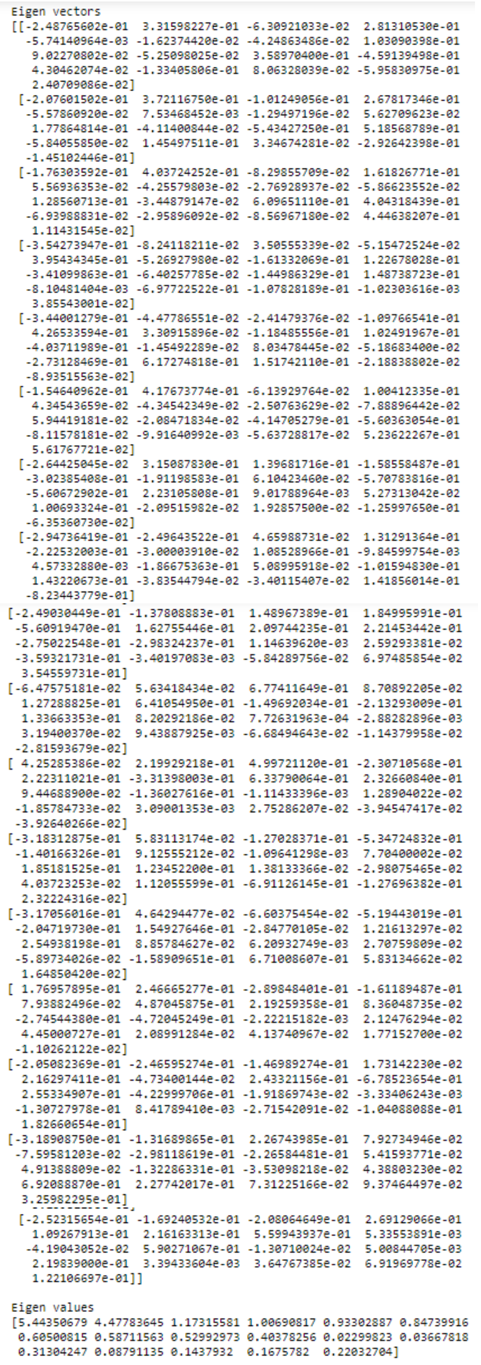
We can observe that there are no changes in outliers before and after scaling.

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

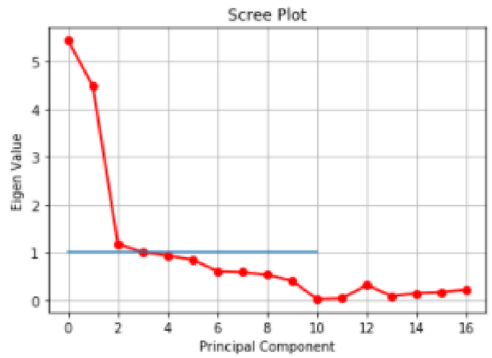


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.176
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.246
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.289
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443	-0.161
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.079
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0.641055	-0.331398	0.091256	0.154928	0.487
6	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-0.149692	0.633790	-0.001096	-0.028477	0.219
7	-0.103090	-0.056271	0.058662	-0.122678	-0.102492	0.078890	0.570784	0.009846	-0.221453	0.213293	-0.232661	-0.077040	-0.012161	-0.083
8	-0.090227	-0.177865	-0.128561	0.341100	0.403712	-0.059442	0.560673	-0.004573	0.275023	-0.133663	-0.094469	-0.185182	-0.254938	0.274
9	0.052510	0.041140	0.034488	0.064026	0.014549	0.020847	-0.223106	0.186675	0.298324	-0.082029	0.136028	-0.123452	-0.088578	0.472
10	0.043046	-0.058406	-0.069399	-0.008105	-0.273128	-0.081158	0.100693	0.143221	-0.359322	0.031940	-0.018578	0.040372	-0.058973	0.445

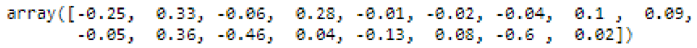
6. Extract the eigenvalues and eigen vectors. [print both]



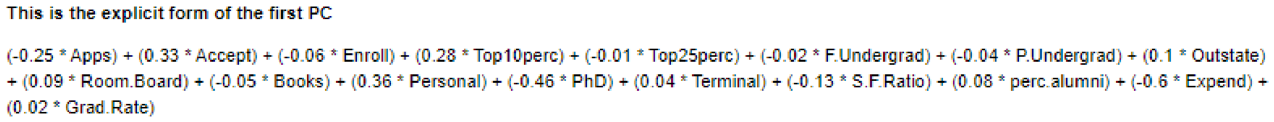
```
Eigen vectors
[[-2.48765602e-01  3.31598227e-01 -6.30921033e-02  2.81310530e-01
  -5.74140964e-03 -1.62374420e-02 -4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-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
 -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]]

Eigen values
[5.44350679 4.47783645 1.17315581 1.00690817 0.93302887 0.84739916
 0.60500815 0.58711563 0.52992973 0.40378256 0.02299823 0.03667818
 0.31304247 0.08791135 0.1437932  0.1675782  0.22032704]
```



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



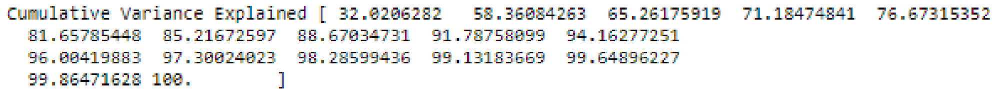
```
array([-0.25,  0.33, -0.06,  0.28, -0.01, -0.02, -0.04,  0.1 ,  0.09,
       -0.05,  0.36, -0.46,  0.04, -0.13,  0.08, -0.6 ,  0.02])
```



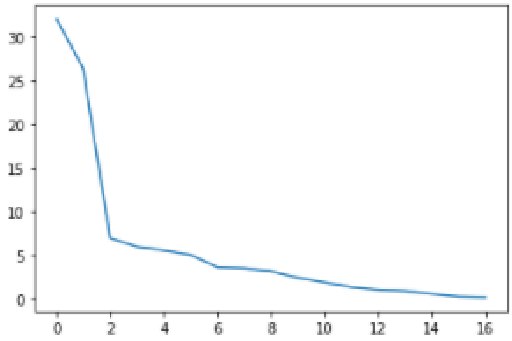
This is the explicit form of the first PC

(-0.25 * Apps) + (0.33 * Accept) + (-0.06 * Enroll) + (0.28 * Top10perc) + (-0.01 * Top25perc) + (-0.02 * F.Undergrad) + (-0.04 * P.Undergrad) + (0.1 * Outstate) + (0.09 * Room.Board) + (-0.05 * Books) + (0.36 * Personal) + (-0.46 * PhD) + (0.04 * Terminal) + (-0.13 * S.F.Ratio) + (0.08 * perc.alumni) + (-0.6 * Expend) + (0.02 * Grad.Rate)

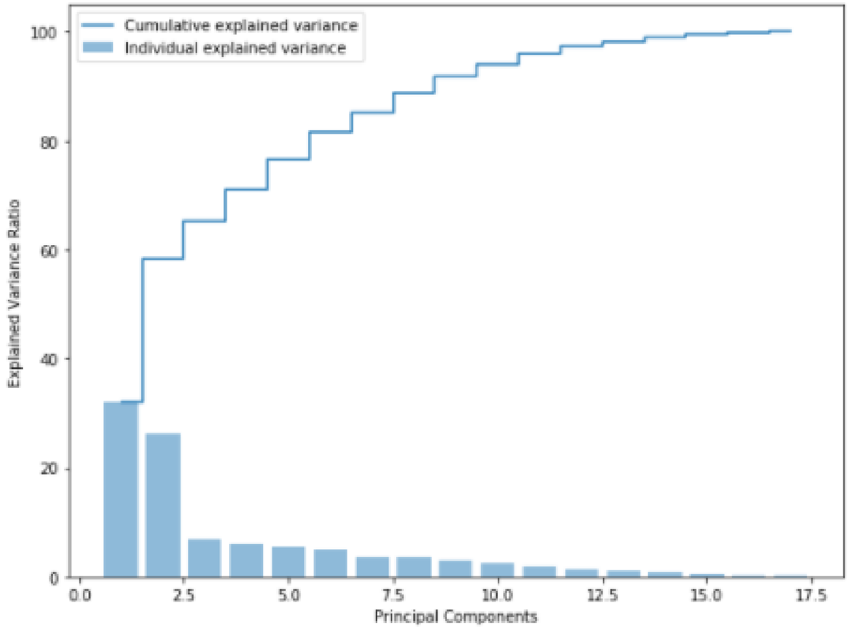
8. **Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?**



```
Cumulative Variance Explained [ 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.        ]
```



- We can observe that the first PC contributes 32% while the second PC contributes 26% and the upcoming PCs contibute lesser and lesser
- All the 17 PCs add upto 100% as indicated in the program above.

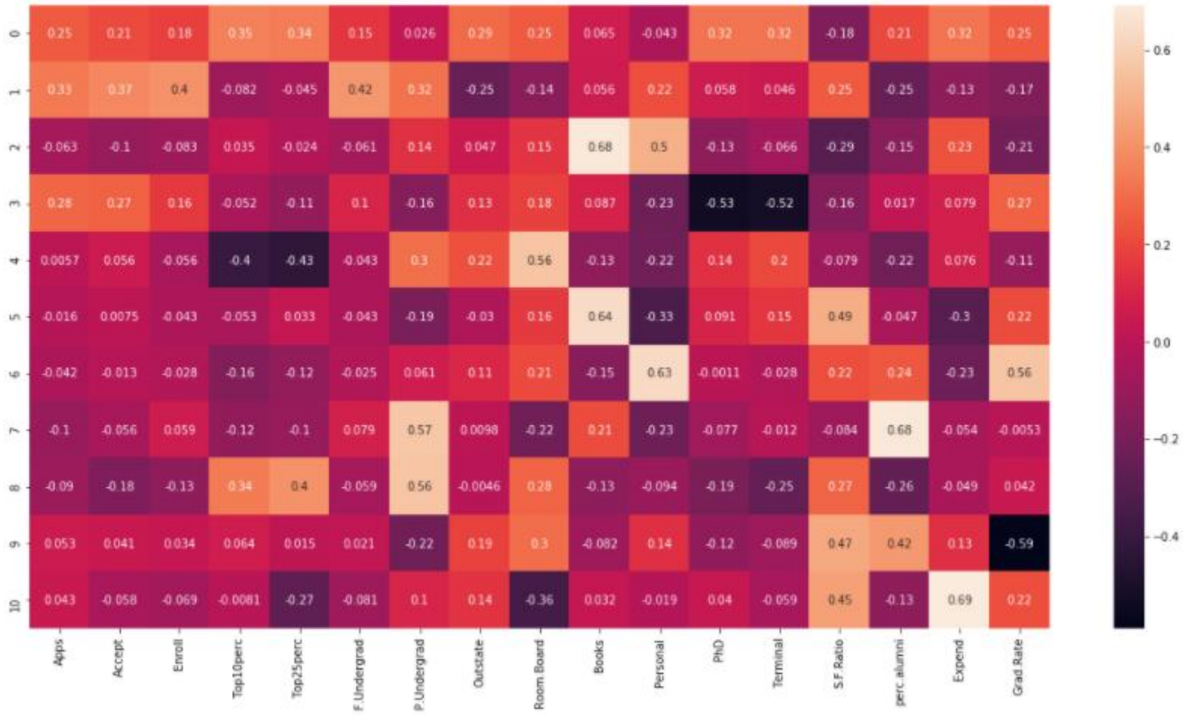


The number of PCs we select is helped by the % of loss of information that is acceptable for this case study.

Eigen values and vectors

- Each Eigen vector is one Principal Component and the next Eigen vector is orthogonal to the previous.
- The corresponding Eigen value determines the strength of the PC.
- Hence it helps us to identify the sequence of the PCs.
- Each element in the Eigen Vector shows us the measure of the corresponding feature variable from the original features list

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]**



	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	0.25	0.21	0.18	0.35	0.34	0.15	0.026	0.29	0.25	0.065	-0.043	0.32	0.32	-0.18	0.21	0.32	0.25
1	0.33	0.37	0.4	-0.082	-0.045	0.42	0.32	-0.25	-0.14	0.056	0.22	0.058	0.046	0.25	-0.25	-0.13	-0.17
2	-0.063	-0.1	-0.083	0.035	-0.024	-0.061	0.14	0.047	0.15	0.68	0.5	-0.13	-0.066	-0.29	-0.15	0.23	-0.21
3	0.28	0.27	0.16	-0.052	-0.11	0.1	-0.16	0.13	0.18	0.087	-0.23	-0.53	-0.52	-0.16	0.017	0.079	0.27
4	0.0057	0.056	-0.056	-0.4	-0.43	-0.043	0.3	0.22	0.56	-0.13	-0.22	0.14	0.2	-0.079	-0.22	0.076	-0.11
5	-0.016	0.0075	-0.043	-0.053	0.033	-0.043	-0.19	-0.03	0.16	0.64	-0.33	0.091	0.15	0.49	-0.047	-0.3	0.22
6	-0.042	-0.013	-0.028	-0.16	-0.12	-0.025	0.061	0.11	0.21	-0.15	0.63	-0.0011	-0.028	0.22	0.24	-0.23	0.56
7	-0.1	-0.056	0.059	-0.12	-0.1	0.079	0.57	0.0098	-0.22	0.21	-0.23	-0.077	-0.012	-0.084	0.68	-0.054	-0.0053
8	-0.09	-0.18	-0.13	0.34	0.4	-0.059	0.56	-0.0046	0.28	-0.13	-0.094	-0.19	-0.25	0.27	-0.26	-0.049	0.042
9	0.053	0.041	0.034	0.064	0.015	0.021	-0.22	0.19	0.3	-0.082	0.14	-0.12	-0.089	0.47	0.42	0.13	-0.59
10	0.043	-0.058	-0.069	-0.0081	-0.27	-0.081	0.1	0.14	-0.36	0.032	-0.019	0.04	-0.059	0.45	-0.13	0.69	0.22

- We observe that with 6 PCs, 81% details of features are captured.
- 80% of the information is carrried by Apps, Accepts, Enroll, Top10perc, Top25perc
- We observe that with 10 PCs, 95% details of features are captured.
- We also observe that only Apps, Accepts & Enroll have eigen values greater than 1.
- Hence, we consider dimensionality reduction from 17 to 3.
- We can remove the other variables from the analysis as they are not as stable as the first 3.
- The dimension of the data is reduced by more than 82%

The number of PCs we select is helped by the % of loss of information that is acceptable for this case study.