ADVANCED STATISTICS PROJECT

Shankar

Report on Salary Data and Education Post 12th Standard using ANOVA Technique and perform Exploratory Data Analysis for Education Post 12th Standard data and perform Principal Component Analysis (PCA) to reduce the dimensions of the data for further Analysis.

PGP – DATA SCIENCE AND BUSINESS ANALYTICS

BATCH: April 2021

DATE: 18/07/2021

TABLE OF CONTENTS

- Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]
- Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?
- Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

PROBLEM 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.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

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

   One Way ANOVA for Education:

   H0: The mean Salary is same across all the 3 categories of Education (Doctorate, Bachelors, HS-Grad)

   H1: The mean Salary is different in at least one Category of Education.

   One Way ANOVA for Occupation:

   H0: The mean salary is same across all the four categories of Occupation (Prof-Specialty, Sales, Adm-Clerical, Exec-Managerial)

   H1: The mean salary is different in at least one category of Occupation.

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.

   One Way ANOVA for Education:

   The below is the ANOVA table for Education. As we can observe that the P-Value for Education (**1.257709e-08**) is much less than the significance level of Alpha (**0.05**) we can **reject** the Null Hypothesis and **accept** Alternate Hypothesis by concluding that there is a significance difference in the mean salaries for at least one category of Education.

	df	sum_sq	mean_sq	F	PR(>F)
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

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.

   One Way ANOVA for Occupation:

   The below is the ANOVA table for Occupation. As we can observe that the P-Value for Occupation (**0.458508**) is greater than the significance level of Alpha (**0.05**) we **fail** to **reject** the Null Hypothesis and conclude that there is no significance difference in the mean salaries across four categories of Occupation.

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

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 Tukey Honest Significant Difference Test is performed to find out which class means are significantly different.

   The Tukey HSD test is performed for Education and the result is displayed in the below table.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7541.1439	79006.9894	True
Bachelors	HS-grad	-90114.1556	0.001	-132035.1958	-48193.1153	True
Doctorate	HS-grad	-133388.2222	0.001	-174815.0876	-91961.3569	True

   As we can observe here, the P-Value for all the three Categories is less than the Alpha. This implies that mean salaries across all different categories of Education is different.

   The Tukey HSD test is performed for Occupation and the result is displayed in the below table.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

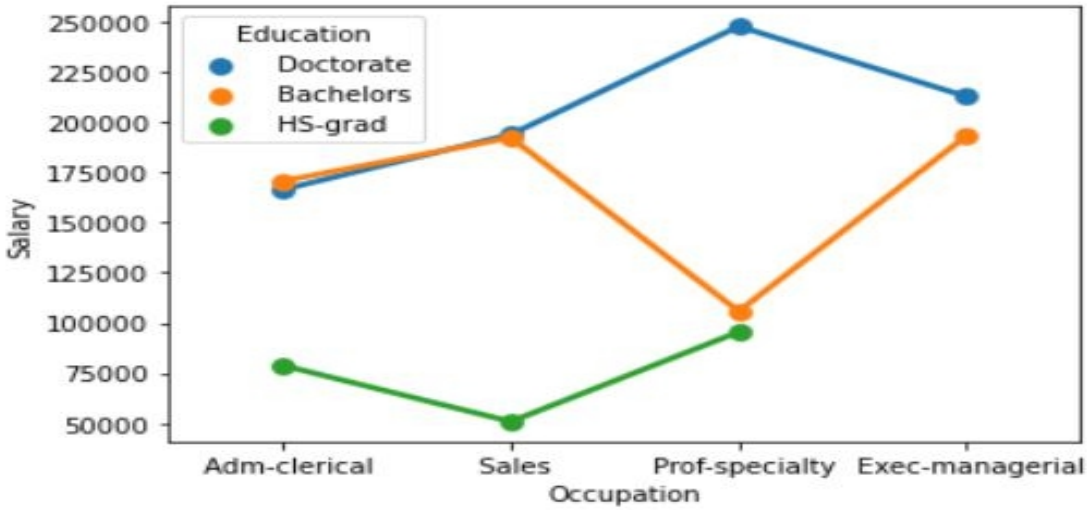
group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

   As we can observe here, the P-Value for all categories of Occupation is higher than Alpha. This implies that the Mean salaries for categories of Occupation are same.

PROBLEM 1B:

1. What is the interaction between two treatments? Analyse 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]

   We can analyse the effects of one variable on the other (Education and Occupation) with the help of an Interaction Plot.



   The interaction plot shows that there is a significant amount of interaction between the two categorical variables (Education and Occupation).

   We can observe a few things by analysing the interaction plot.

   - People with HS-Grad as Education do not reach the level of Exec-Managerial as Occupation.
   - People with Education as Bachelors or Doctorate and Occupation as Sales and Adm-Clerical almost have the same salary.
   - People with Bachelors as Education and Occupation as Prof-Specialty earn less than others with Bachelors as Education.
   - A person with Education as Doctorate and Occupation as Prof-Specialty earns the highest Salary.
   - People with HS-Grad as Education have the lowest range of Salaries. (50000 – 100000)

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?

   Two Way ANOVA (Education and Occupation)

   H0: The effect of independent variable Education on the mean Salary does not depend on the effect of the other independent variable Occupation.

   H1: There is interaction between the two independent variable Education and Occupation with respect to mean Salary.

   After performing Two Way ANOVA technique to both Education and Occupation with the mean salary, we get the below table.

	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

   From the table we can observe that the P-Value for Education and Occupation is less than the significance level of Alpha. But eventually we need to understand the interaction between the two independent variables.

   After performing the Two Way ANOVA for Education and Occupation considering their interaction we get the below table.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

From this we can understand there is interaction between the two independent variables Education and Occupation.

The P-Value of Education and Occupation (**2.232500e-05**) is less than the significance level of Alpha (**0.05**). We can reject the Null hypothesis and accept the Alternate Hypothesis by concluding that there is interaction between the two variables *Education* and *Occupation* on mean *Salary*.

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

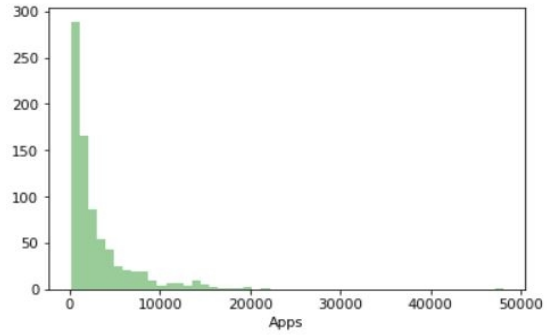
   From the ANOVA method and the interaction plot we can conclude that Education combined with Occupation results in higher Salaries. It is clearly seen that the people with Education as Doctorate draw maximum amount of Salaries where as people with HS-Grad as Education Background earns the least amount of Salaries.

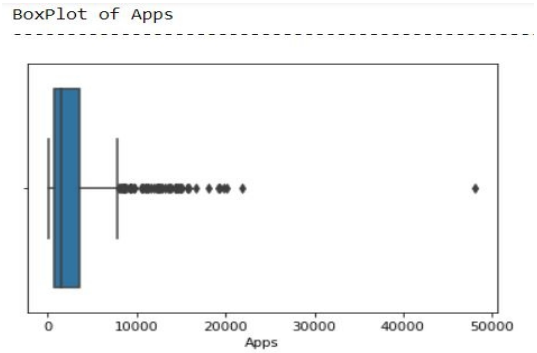
   Thus we can conclude that **Salary** is dependent on both "**Educational Qualifications** and **Occupation"**.

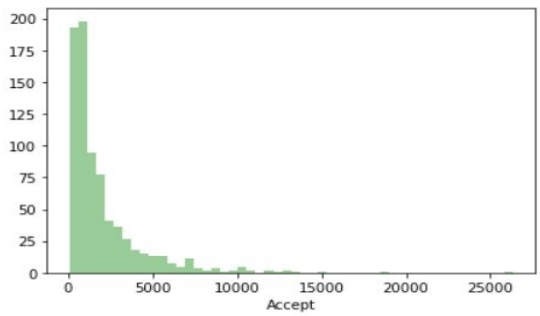
PROBLEM 2:

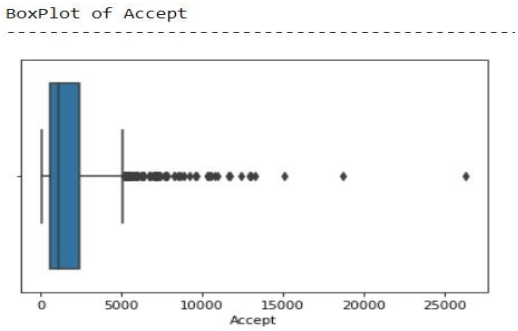
The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data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

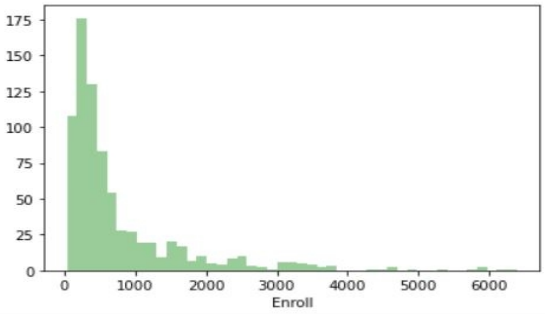
- Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

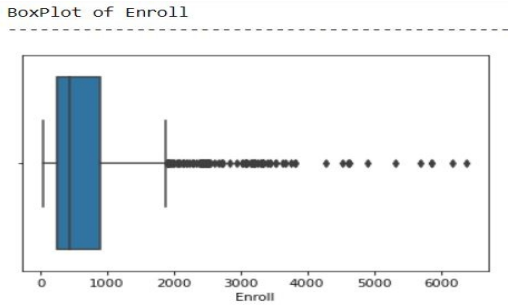


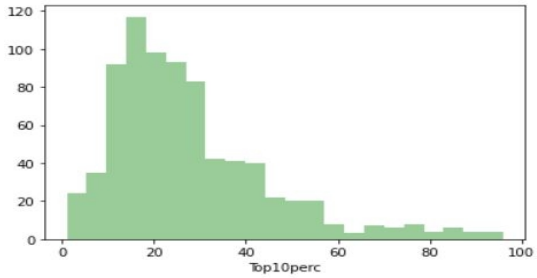


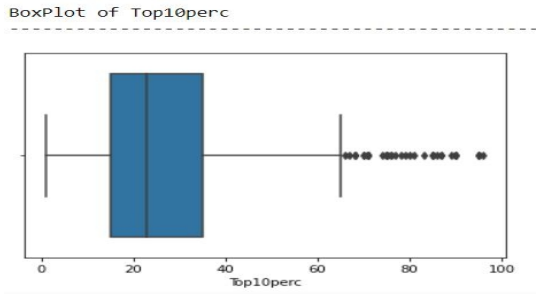


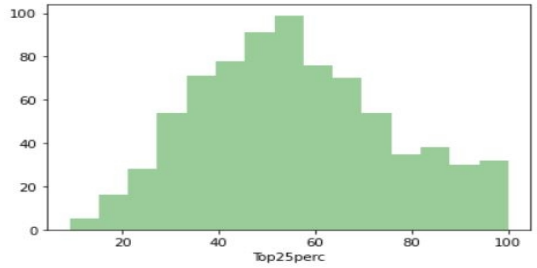


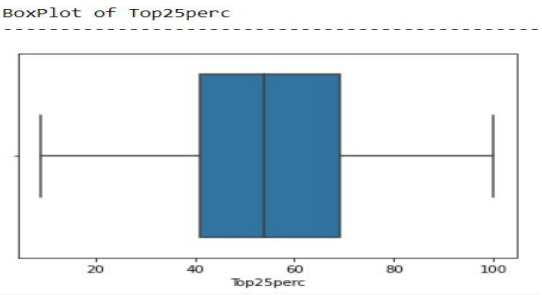


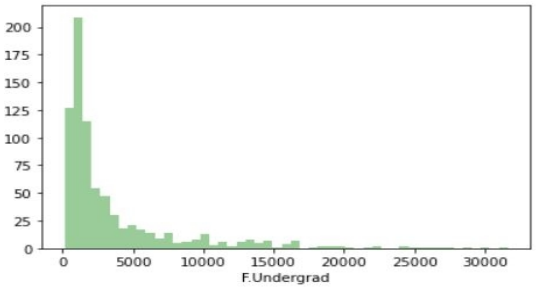


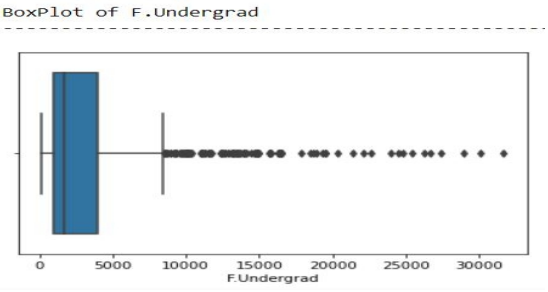


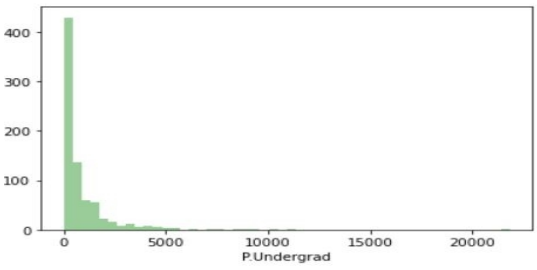


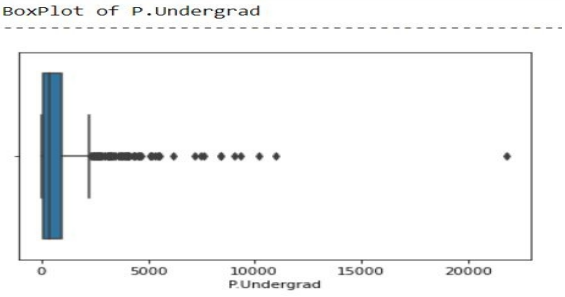


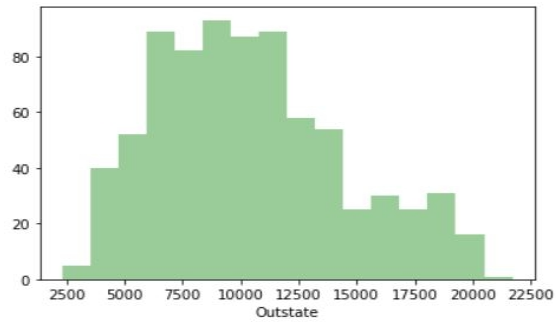


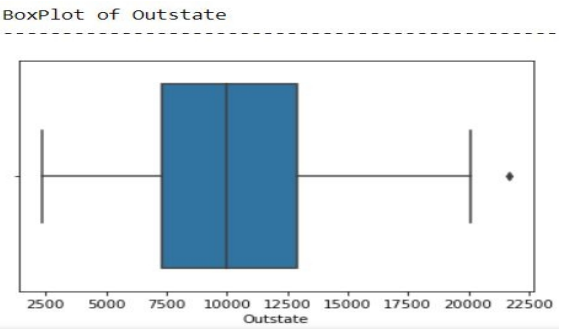


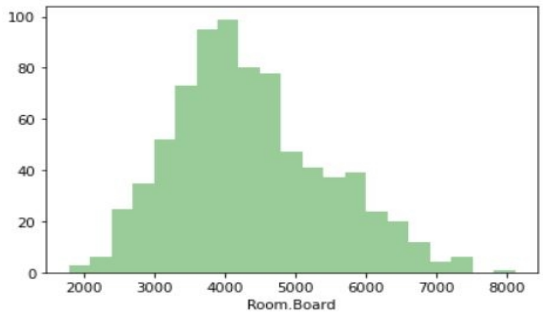


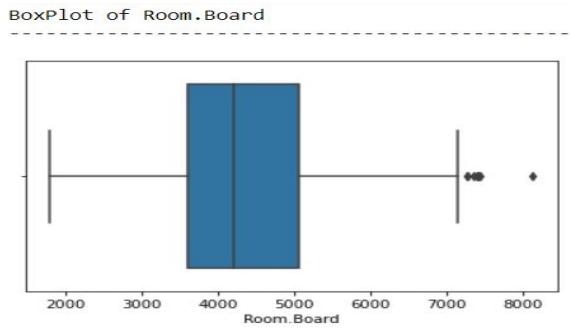


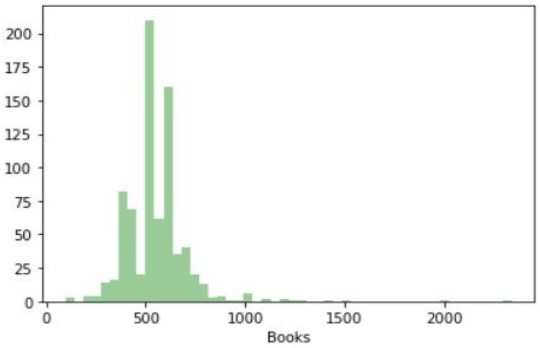


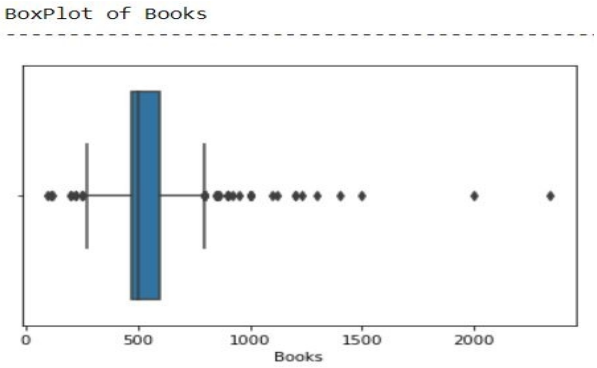


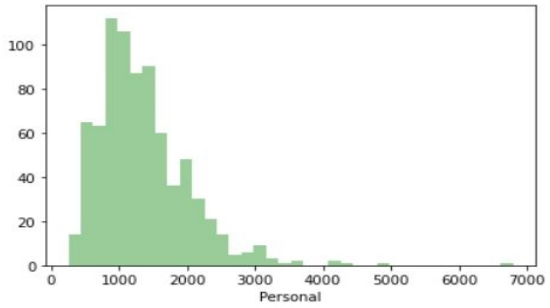


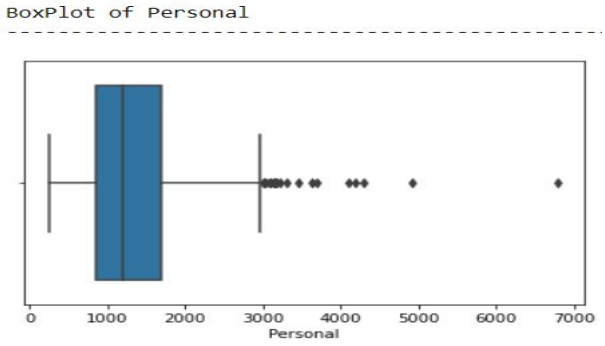


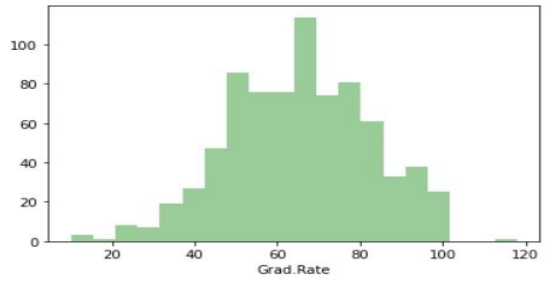


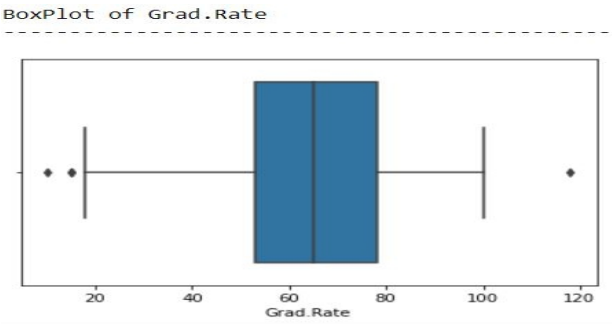




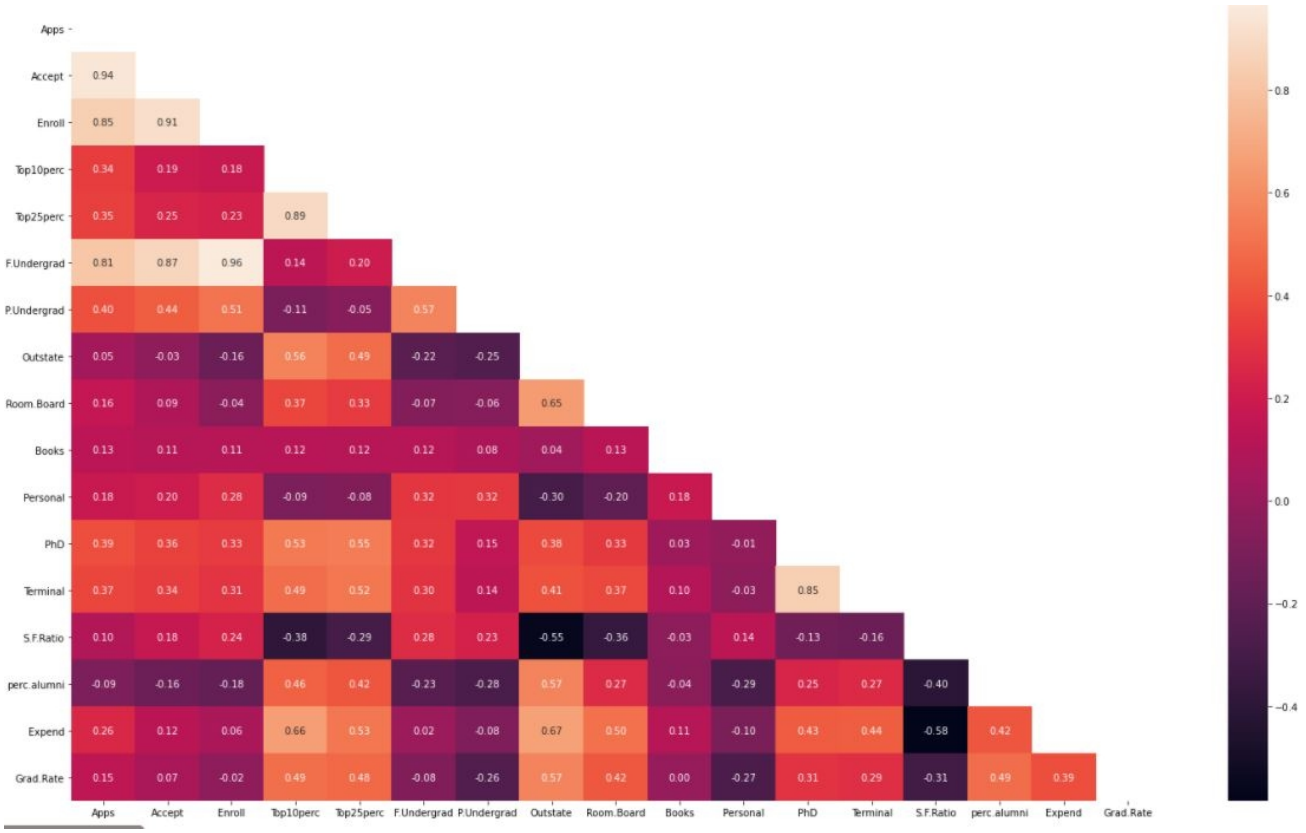








These are some of Univariate Analysis of the given Data set. The Multivariate Analysis is shown using the Pearson Heat Map which also helps us to derive the Correlation between the fields.



OBSERVATIONS:

- There are 17 numeric fields in the data
- The number of application received ranges from 81 to 48094
- The number of application accepted ranges from 72 to 26330
- The number of application enrolled ranges from 35 to 6392
- The percentage of new students accepted is more from top 25% (139 to 31643)
- Most of the student prefers to go for F.Undergrad compared to P.Undergrad
- The cost of Room and Board ranges from 1780 to 8124 with an average of 4200
- Average percentage of faculties with PhD is 75%
- About 78% students are graduating
- Outliers to be treated
- We can able to see a lot fields have a strong correlation.
- Apps shows high correlation with Accept, Enroll and F.Undergrad

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

  Yes Scaling is necessary for PCA in this given Case Study.
  - We can observe that we have fields where the ranges are different from each other.
  - Scaling is performed where the Dataset has different features with different weights.
  - So we perform Scaling so that all the features in the Dataset are on the same '*Scale*'.
  - In our Dataset we can observe that we have fields with APPS where it ranges from 81 to 48094 and we have fields with Top10Perc where it ranges from 1 to 35.
  - So we have Scale our data in order to further analyse and visualize it to gather more information from the given Dataset.

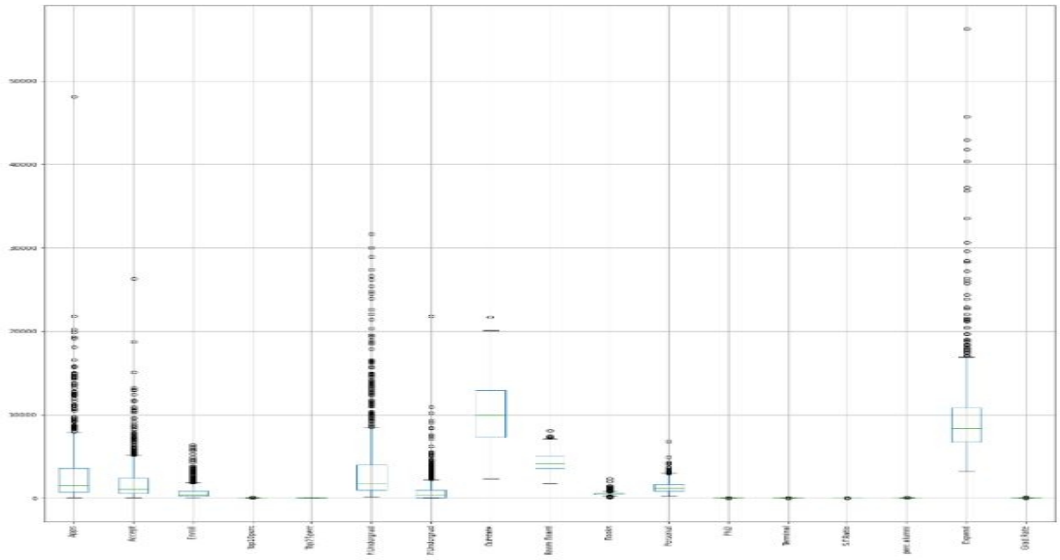
- Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
  - Using the correlation matrix is equivalent to standardizing each of the variables (to mean 0 and standard deviation 1). In general, PCA with and without standardizing will give different results. Especially when the scales are different.
  - We tend to use the Covariance Matrix when the variable Scales are similar and the correlation matrix when variables are on different Scales.

	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	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.095756	-0.090342	0.259927	0.146944
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338018	0.176456	-0.160196	0.124878	0.067399
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308671	0.237577	-0.181027	0.064252	-0.022370
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491768	-0.385370	0.456072	0.661765	0.495627
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525425	-0.295009	0.418403	0.528127	0.477896
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300406	0.280064	-0.229758	0.018676	-0.078875
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142086	0.232830	-0.281154	-0.083676	-0.257332
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408509	-0.555536	0.566992	0.673646	0.572026
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375022	-0.363095	0.272714	0.502386	0.425489
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100084	-0.031970	-0.040260	0.112554	0.001062
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950	-0.030653	0.136521	-0.286337	-0.098018	-0.269691
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289	0.850682	-0.130698	0.249330	0.433319	0.305431
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682	1.001289	-0.160310	0.267475	0.439365	0.289900
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.160310	1.001289	-0.403448	-0.584584	-0.307106
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330	0.267475	-0.403448	1.001289	0.418250	0.491530
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319	0.439365	-0.584584	0.418250	1.001289	0.390846
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431	0.289900	-0.307106	0.491530	0.390846	1.001289

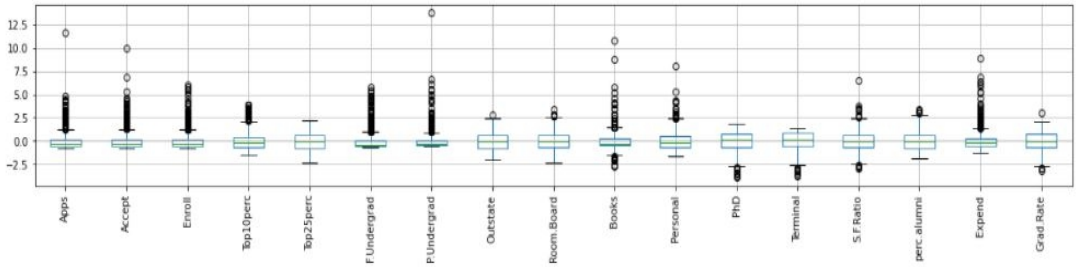
As we can see the Covariance Matrix for the given Dataset after Scaling, we can observe that a lot of features have a strong correlation between each other. On the scaled data the correlation matrix describes the strength of each field with the others.

- 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]

  We can observe the most number of fields on the given dataset have outliers and on a different scales. This is shown in the below graph.



  After performing the Scaling for the given Dataset and plotting the graph we get



- After performing the scaling for the given dataset we can observe that the outlier of the given Dataset ranges from -2.5 to 12.5 whereas before scaling the Dataset ranges from 0 to 90000 which is a larger range.
- After performing Scaling we can use the Dataset to perform PCA which will enable us to acquire more accurate result.
- By performing Scaling we are making all the fields in the given Dataset to have Single Scale which will help us in further analysis.

- Extract the eigenvalues and eigenvectors.

  The Eigen Values and Eigen Vectors of the given Dataset without reducing the dimensions are

```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01, [-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,  -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,  -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,   6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,   1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
         3.18908750e-01,  2.52315654e-01],                  -2.98118619e-01,  2.16163313e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01, [-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,  -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,   6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,  -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,  -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
        -1.31689865e-01, -1.69240532e-01],                 -2.26584481e-01,  5.59943937e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02, [-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,  -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,   5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,   2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,  -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
         2.26743985e-01, -2.08064649e-01],                  -5.41593771e-02, -5.33553891e-03],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01, [-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,   3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,   5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,  -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,  -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
         7.92734946e-02,  2.69129066e-01],                  -4.91388809e-02,  4.19043052e-02],
       [ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02, [ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
        -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,   6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
         3.02385408e-01,  2.22532003e-01,  5.60919470e-01,  -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
        -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,  -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
         2.04719730e-01, -7.93882496e-02, -2.16297411e-01,  -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
         7.59581203e-02, -1.09267913e-01],                   1.32286331e-01, -5.90271067e-01],
```

```
[ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
 -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
  1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
  3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
 -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
  6.92088870e-01,  2.19839000e-01],
[ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
  3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
 -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
 -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
  1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
  3.25982295e-01,  1.22106697e-01],
[ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
  1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
  1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
  1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
 -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
 -9.37464497e-02, -6.91969778e-02],
[ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
 -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
  1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
 -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
  6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
  7.31225166e-02,  3.64767385e-02],
[ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
  6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
  2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
 -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
  1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
 -2.27742017e-01, -3.39433604e-03],
[ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
 -1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
 -5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
  2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
 -2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
 -4.38803230e-02, -5.00844705e-03],
[ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
 -1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
  9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
  7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
  6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
 -3.53098218e-02, -1.30710024e-02]])
```

The Eigen Values are

```
array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871, 0.03453621, 0.03117234, 0.02375192,
       0.01841426, 0.01296041, 0.00985754, 0.00845842, 0.00517126,
       0.00215754, 0.00135284])
```

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

  - To perform PCA we have to conduct two tests which are "**Bartletts Test of Sphericity and KMO Tests".**

  - The results of the **Bartletts Test of Sphericity** are **0.0** and **KMO Tests** is **0.8184659398241376.**

  - The KMO tests results approves that the dimensions of the given dataset can be reduced since the value is greater than 0.75 (which states that there is chance to reduce the number of dimensions).

- After performing PCA the data of the Principal Component are exported to **df_pca**.

- Principal Component Analysis is performed for the given dataset and the dimension has been reduced to 8. The Eigen Vectors of the Data Frame after PCA are

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.0630921 , -0.10124906, -0.08298557,  0.03505553, -0.02414794,
        -0.06139298,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131053,  0.26781735,  0.16182677, -0.05154725, -0.10976654,
         0.10041234, -0.15855849,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.07927349,  0.26912907],
       [ 0.00574141,  0.05578609, -0.05569364, -0.39543434, -0.42653359,
        -0.04345437,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
        -0.22231102,  0.14016633,  0.20471973, -0.07938825, -0.21629741,
         0.07595812, -0.10926791],
       [-0.01623744,  0.00753468, -0.04255798, -0.0526928 ,  0.03309159,
        -0.04345423, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
        -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
        -0.29811862,  0.21616331],
       [-0.04248635, -0.01294972, -0.02769289, -0.16133207, -0.11848556,
        -0.02507636,  0.06104235,  0.10852897,  0.20974423, -0.14969203,
         0.63379006, -0.00109641, -0.02847701,  0.21925936,  0.24332116,
        -0.22658448,  0.55994394],
       [-0.1030904 , -0.05627096,  0.05866236, -0.12267803, -0.10249197,
         0.07888964,  0.57078382,  0.009846  , -0.22145344,  0.21329301,
        -0.23266084, -0.07704   , -0.01216133, -0.08360487,  0.67852365,
        -0.05415938, -0.00533554]])
```

The Coefficients of the Eight Principal Components are

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Apps	0.25	0.33	-0.06	0.28	0.01	-0.02	-0.04	-0.10
Accept	0.21	0.37	-0.10	0.27	0.06	0.01	-0.01	-0.06
Enroll	0.18	0.40	-0.08	0.16	-0.06	-0.04	-0.03	0.06
Top10perc	0.35	-0.08	0.04	-0.05	-0.40	-0.05	-0.16	-0.12
Top25perc	0.34	-0.04	-0.02	-0.11	-0.43	0.03	-0.12	-0.10
F.Undergrad	0.15	0.42	-0.06	0.10	-0.04	-0.04	-0.03	0.08
P.Undergrad	0.03	0.32	0.14	-0.16	0.30	-0.19	0.06	0.57
Outstate	0.29	-0.25	0.05	0.13	0.22	-0.03	0.11	0.01
Room.Board	0.25	-0.14	0.15	0.18	0.56	0.16	0.21	-0.22
Books	0.06	0.06	0.68	0.09	-0.13	0.64	-0.15	0.21
Personal	-0.04	0.22	0.50	-0.23	-0.22	-0.33	0.63	-0.23
PhD	0.32	0.06	-0.13	-0.53	0.14	0.09	-0.00	-0.08
Terminal	0.32	0.05	-0.07	-0.52	0.20	0.15	-0.03	-0.01
S.F.Ratio	-0.18	0.25	-0.29	-0.16	-0.08	0.49	0.22	-0.08
perc.alumni	0.21	-0.25	-0.15	0.02	-0.22	-0.05	0.24	0.68
Expend	0.32	-0.13	0.23	0.08	0.08	-0.30	-0.23	-0.05
Grad.Rate	0.25	-0.17	-0.21	0.27	-0.11	0.22	0.56	-0.01

- Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

	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	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.176958	0.205082	0.318909	0.252316

The First Principal Component has been displayed along with its features.

The Eigen Vector of first PC is

```
[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
  0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
 -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
  0.31890875,  0.25231565],
```

- 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?

  The Eigen Values of the given Dataset are

```
array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871, 0.03453621])
```

  We can observe that the sum of these 8 Eigen Values is **0.85216725.**

  These 8 Principal Components is Contributing 85.22% of the total variance. Thus we can reduce the dimensions to 8 from 33 so that further analysis can be done.

  The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude. In other words, the eigenvalues explain the variance of the data along the new feature axes.

  These eight Eigen Vectors the indicated the vectors who values is not affected by the **Linear Transformation**. These eight eigenvectors indicates the vectors whose direction is not affected after the Scaling the data.

- 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]

  After using the Principal Component Analysis for the given data, we have reduced the number of dimensions from the main Dataset.

  This can help us to calculate the percentage who got acceptance letter from the University have enrolled and graduated from the University. This helps to identify the Best University for the Students.

Also this can also be used for the students to classify the best University in his home town rather than going Outstate.

For further Analysis we can have "**Eight Principal Components**" which is sufficient rather than work with the main Dataset.

The correlation of the Eight Principal Components is shown in the below diagram.

By seeing this we can that the Principal Components are strongly correlated with few features. These Eight Principal Components can be used for further analysis to get more accurate result and to pin point the Best Universities for Students to apply in Outstate or in hometown.

END