Advanced Statistics PROJECT REPORT

DSBA

By- Aarti Londhe

greatlearning
Learning for Life

Contents

Table of Contents

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.
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.
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.
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)**
5. 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]
6. 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?
7. Explain the business implications of performing ANOVA for this particular case study.

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

Salary is hypothesized to depend on educational qualification and occupation.

In order to conduct One-Way ANOVA below are our null and alternate hypothesis.

For Education,

Null Hypothesis is that mean salary of persons is independent of their Education and is observed to be equal for all Education level.

Alternate Hypothesis is that mean salary is different for all education level or for at least one Education Level.

H0 : The means of 'Salary' variable with respect to each Education level is equal.

H1 : At least one of the means of 'Salary' variable with respect to each Education level is unequal.

For Occupation,

Null Hypothesis is that mean salary of persons is independent of their Occupation and is observed to be equal for all Occupation level.

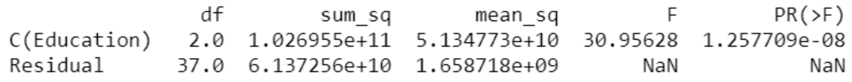
Alternate Hypothesis is that mean salary is different for all occupation level or for at least one Occupation Level.

H0 : The means of 'Salary' varies with respect to each Occupation level is equal.

H1 : At least one of the means of 'Salary' varies with respect to each Occupation level is unequal.

1.2. 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

Based on One-Way ANOVA test on Salary with respect to Education,



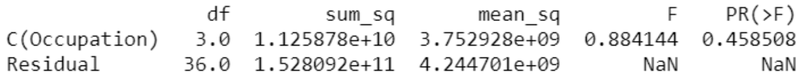
```
                 df        sum_sq       mean_sq         F        PR(>F)
C(Education)    2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual       37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

Based on above ANOVA result , calculated F statistics value of 1.257709e-08 is less than alpha 0.05 (Confidence level), we reject the NULL hypothesis.

i.e. population means for salary are not same for difference Education levels. In other words Salary is affected by Education of person and is a significant factor for salary of person.

1.3. 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.

Based on One-Way ANOVA test on Salary with respect to Occupation,



```
                  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
```

Based on above ANOVA result , calculated F statistics value of 0.4585 is greater than alpha 0.05 (Confidence level), we failed to reject the NULL hypothesis.

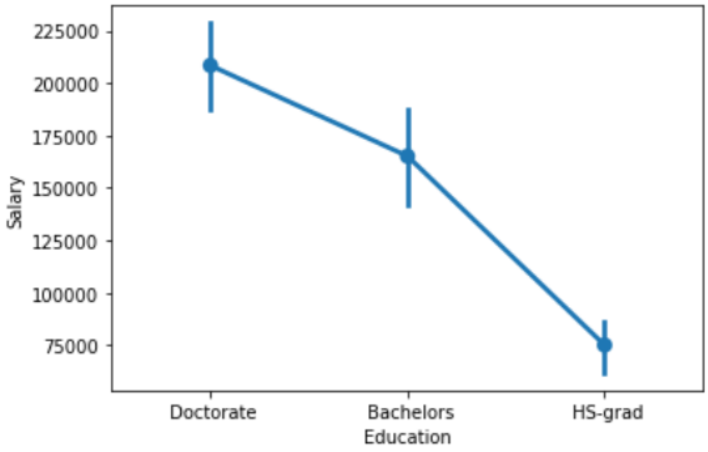
i.e. population means for salary are same for difference Occupation levels. In other words Salary is not affected by Occupation of person and is not a significant factor for salary of person.

1.4. 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)

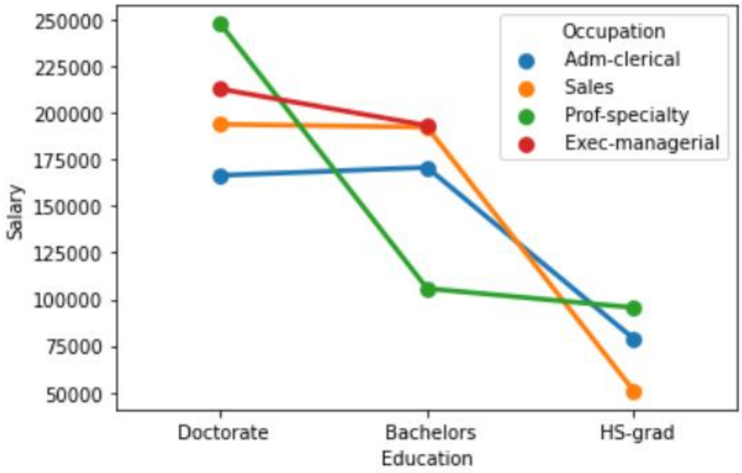
Null Hypothesis is rejected in 1.2
i.e. Mean salary of person is affected by different Education levels.

In order to interpret the class means for each education level, we can use pointplot.

Below graph displays mean salary for difference Education class.
Education Class- Doctorate has significantly large mean salary compared to HS-grad Education mean salary.



1.5. 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]



We have observed that for each Education level there is a different mean salary which is further impacted or calibrated by different occupation levels.

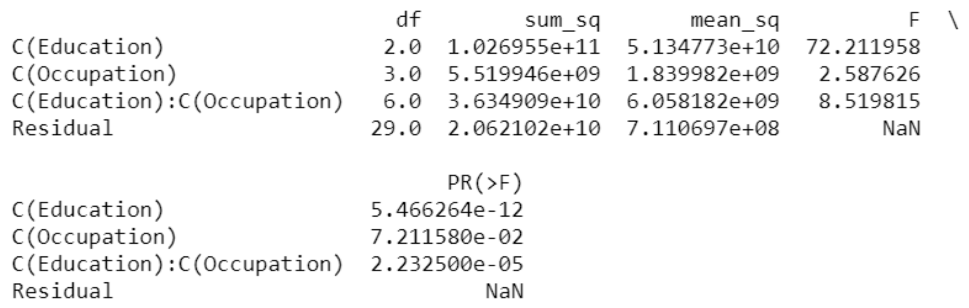
That is there is a interaction between Occupation and Education level which has an effect on Mean Salary of person.

1.6. 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 is that there is not a interaction between two treatments and no effect on mean salary.
Alternate Hypothesis is that there is a interaction between two treatments.

H0 : The means of 'Salary' variable with respect to each Education and Occupation level is equal.
H1 : At least one of the means of 'Salary' variable with respect to each Education and Occupation level is unequal.

Two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). -



```
                              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
```

Here from above anova table, F value 2.232500e-05 is less than 0.05 Alpha
hence null hypothesis is rejected .

We have enough evidence that there is an interaction between two treatments, and it affect the mean salary of person.

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

Above ANOVA technique has helped to understand cause and effect relation among the independent variable Education and Occupation on Dependent variable Salary.

By one way anova , we can say that Education level have a significant impact on mean salary of person i.e. it is a significant cause of effect on mean salary of a person, but there is not enough evidence that Occupation has significant impact on mean salary of person.

Also using a two way ANOVA, we have observed that, there is a significant impact of interaction between Education and occupation level on mean salary of person.

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.

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

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

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

2.4. 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]

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

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

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

2.8. 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?

2.9. 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]

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

In Given Data set, there are 17numeric fields and 1 categorical field i.e. Names: Names of various university and colleges.

In attached python file we have performed univariate analysis on dataset df using a function 'univariateAnalysis_numeric' which display information as part of univariate analysis of numeric variables. The function will accept column name.

The function will display the statistical description of the numeric variable, distplot to view the distribution and the box plot to view 5 point summary and outliers if any.

Observations from Univariate analysis:

- There are 17 numeric fields in the data.
- On an average 3000 applications are received per colleges for Post 12th Standard Education.
- Average number of Full time under-graduate students is more compared to part-time under graduate students.
- Cost of Room and Board varies from 1780 Rs to 8124 Rs. (..considering currency as Rs)
- Average Estimated cost of Books per student is 549RS
- Estimated personal spending per student varies from 250Rs to 6800Rs.Average personal spending per student is 1340Rs.
- On an average 72% faculties have Ph.D degree while 79% faculties have terminal degree.
- Avg. Student to faculty ratio for given list of colleges is 14%.
- Its observed that around 22% of alumni donate
- The average Instructional expenditure per student is of 9660Rs.
- From past data, average Graduation rate is observed to be 65%.

In attached python file we have performed univariate analysis on dataset df using a pairplot and heatmap which display correlation among variables of dataset.

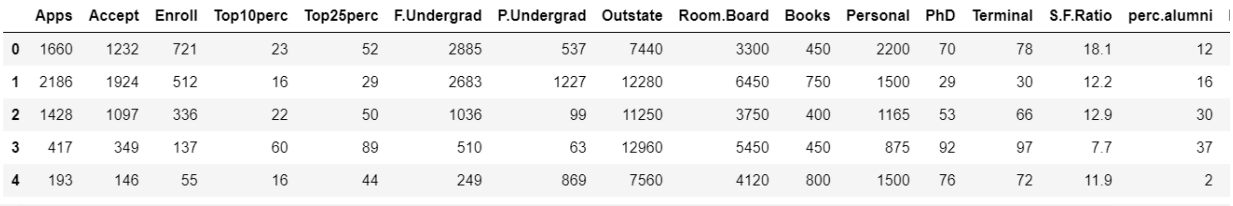
- All variables seems to have linear relation with each other.
- Out of the total new students, maximum has enrolled for Full Time Undergraduation.
- The Instructional expenditure per student is observed to be more for students to whom the particular college or university is Out-of-state tuition

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

- Scaling is necessary for PCA.
- PCA effectiveness depends upon the scales of the attributes. If attributes have
- different scales, PCA will pick variable with highest variance rather than picking up
- attributes based on correlation. So, it is important that all data attributes should be converted to same scale.
- Changing scales of the variables can change the PCA.

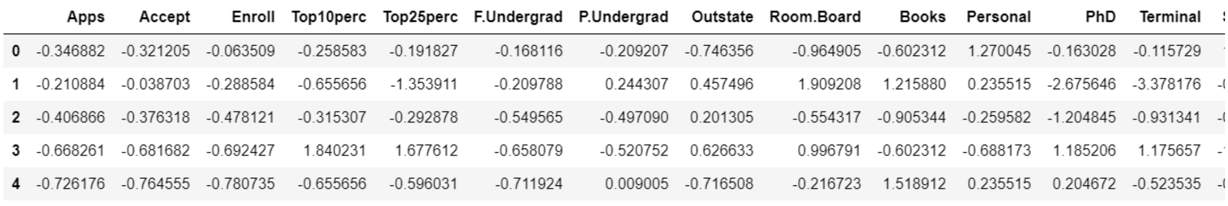
To perform scaling in python, we have used ZSCORE method.

<u>Data before scaling:</u>



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni
0	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12
1	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16
2	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30
3	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37
4	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2

<u>Data after scaling:</u>



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535

- Scaled data shows range for all attribute values in similar range.
- Scaling centralizes the data to origin as it minus the data values from it mean.

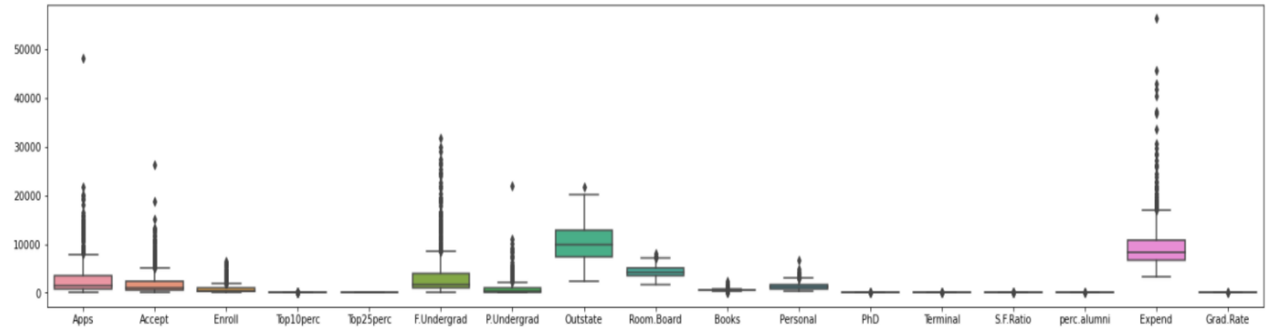
2.3. Comment on the comparison between the covariance and the correlation matrices from this data.

- Here we have performed covariance on scaled numeric data.
- From attached results in python for covariance and correlation matrix, we can say both the matrices are same. For both Matrices output values varies from -1 to +1.
- Covariance matrix of scaled data is same as correlation matrix of the given data.
- Covariance value displays the variance within the variable and covariance among the variables. It measures the direction of linear relationship between the variables.
- In case of scaled data, Data on all the dimensions are subtracted from their means to shift the data points to the origin. i.e. the data is centered on the origins.
- So, covariance matrix of scaled data is same as correlation matrix of data.

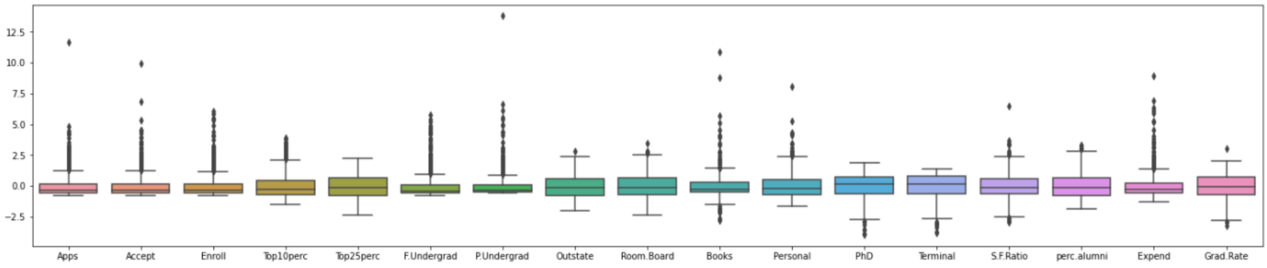
2.4. Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so].

We can use Boxplot to check outliers.

Boxplot before scaling of the data:



Boxplot after scaling of the data:



- Outliers are still observed for the data after scaling.

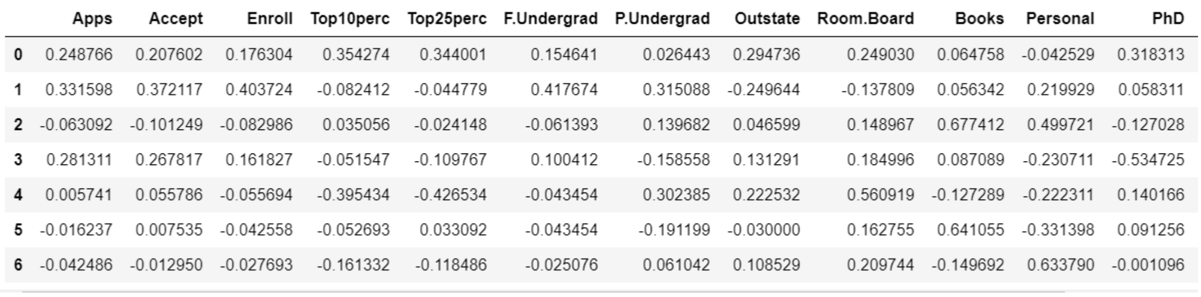
2.5. Perform PCA and export the data of the Principal Component scores into a data frame.

We can perform PCA using a library function from sklearn.decomposition.

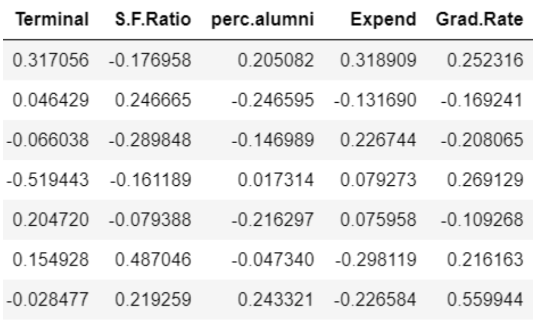
Here we will give input number of compoenents as 7 i.e. we are generating only 7 PCA dimensions (dimensionality reduction from 17 to 7).

Below represents the 7x17 matrix of PCA Componenets.

Dataframe of PCA components.



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



Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0.317056	-0.176958	0.205082	0.318909	0.252316
0.046429	0.246665	-0.246595	-0.131690	-0.169241
-0.066038	-0.289848	-0.146989	0.226744	-0.208065
-0.519443	-0.161189	0.017314	0.079273	0.269129
0.204720	-0.079388	-0.216297	0.075958	-0.109268
0.154928	0.487046	-0.047340	-0.298119	0.216163
-0.028477	0.219259	0.243321	-0.226584	0.559944

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

Eigen Vectors

```
%s [[-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

```
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
 0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
 0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

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

The explicit form of the first PC:

0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * F.Undergrad + 0.03 * P.Undergrad + 0.29 * Outstate + 0.25 * Room.Board + 0.06 * Books + -0.04 * Personal + 0.32 * PhD + 0.32 * Terminal + -0.18 * S.F.Ratio + 0.21 * perc.alumni + 0.32 * Expend + 0.25 * Grad.Rate

2.8. 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 Values

```
[ 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.        ]
```

From above Cumulative values of Eigen values , we can say that first principle component captures 32% of variance , first Two PC captures 58% , first three PC captures 65%, first four PC captures 71%, first five PC captures 76% and so on.

If we consider first 7 compoenent as principle component, we can say that it will captures the variance of 85% of the data.

Eigen vectors are the coefficients of principle component which together with our original variables/features decides the first principle component dimension. Eigen vector decides the direction of maximum varaince of the data.

2.9. 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]

With help of PCA we have been able to reduce 17 numeric features into 7 components which is able to explain 85% of variance in the data. It can also help to understand which factors affects the maximum variance of the data.

Principle components reduces the linear relation between variables.

With Eigen Vector on Covaraince matrix, we identified the direction which captures the maximum variance of the data by reducing the correlation among the variables.

Eigen values calculation helped to decide the number pf PCs in terms of their variance calculation.