# Advanced Statistics - Report

## Contents

2.7 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]

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]

# Problem 1A

**Problem Statement:**

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.

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

**1A.1 Answer:**

**Education:**

- H0: Salary depend on education qualification
- H1: Salary does not depend on education
- Confidence level = 0.05

**Occupation:**

- H0: Salary depend on occupation
- H1: Salary does not depend on occupation
- Confidence level = 0.05

1A.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.

**1A.2 Answer:**

- F Statistic is 30.956 and P is 0.0000000126
- Since P Value is less than 0.05 we **reject Null Hypothesis**

1A.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.

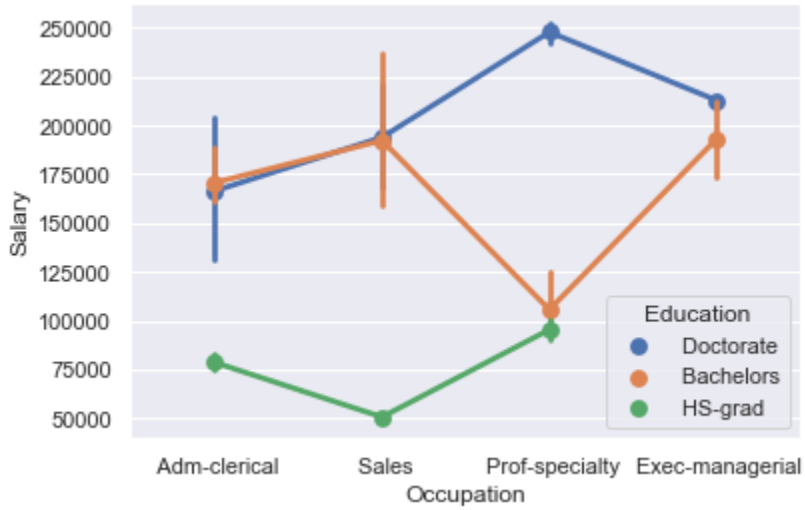
**1A.3 Answer:**

- F Statistic is 0.884 and P is 0.4585078266
- Inference: Since P Value is less than 0.05 we **reject Null Hypothesis**

# Problem 1B

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]

**1B.1 Answer:**




**Inference**: Adm-clerical and Sales people with Bachelors and Doctorate Degrees earn almost similar salary packages

1B.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?

**1A.2 Answer:**

- H0: Salary depends on both Education and Occupation
- H1: Salary does not depend on at least one of Education and Occupation
- Confidence level = 0.05

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **C(Education)** | 1.938753e+11 | 2.0 | 136.326521 | 1.756909e-12 |
| **C(Occupation)** | 4.077417e+08 | 3.0 | 0.191140 | 8.270491e-01 |
| **C(Education):C(Occupation)** | 4.227791e+10 | 6.0 | 9.909463 | 1.323371e-05 |
| **Residual** | 2.062102e+10 | 29.0 | NaN | NaN |

**Inference**: Education is a significant factor as the P value is <0.05, whereas Occupation is not a significant variable as P value of it is >0.05

1B.3 Explain the business implications of performing ANOVA for this particular case study.

**1B.3 Answer:**

It can be concluded that Education is a significant factor.

Salary is dependent on occupation.

# Problem 2

**Problem Statement:**

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?

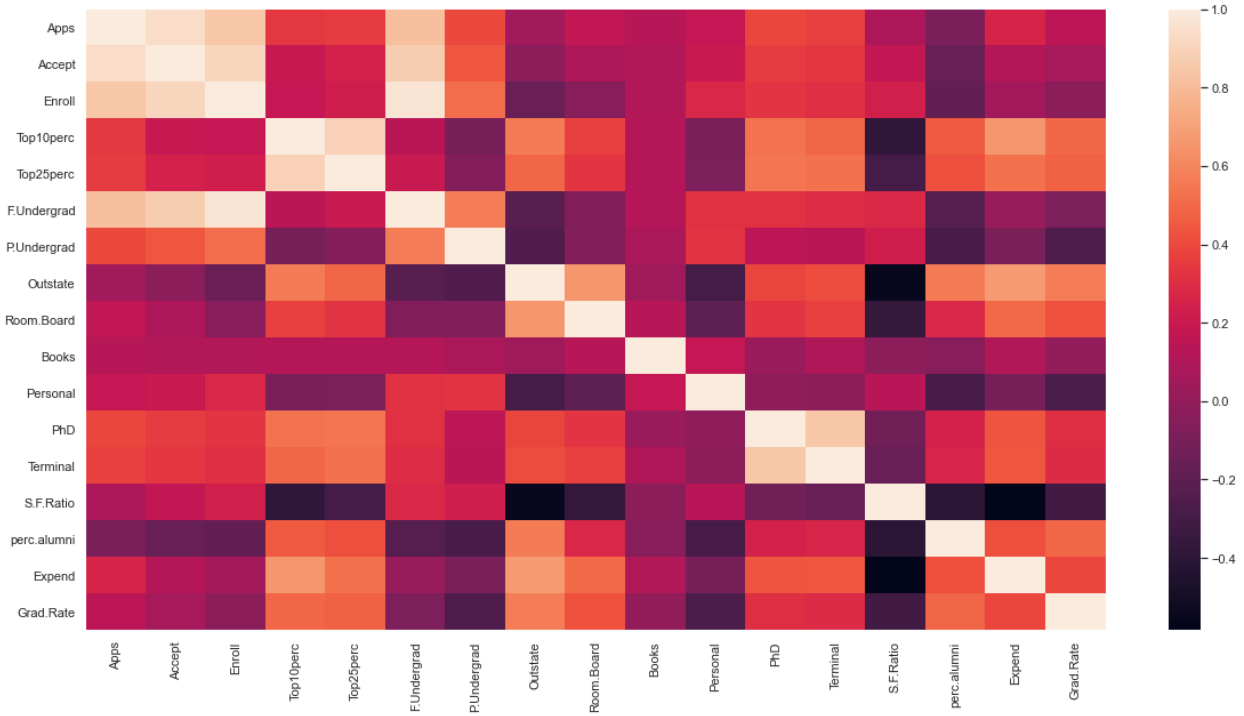
**2.1 Answer:**

```
df_3.isnull().sum()
```

```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

```
df_3.describe()
```

| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 77 |
| **mean** | 3001.638353 | 2018.804376 | 779.972973 | 27.558559 | 55.796654 | 3699.907336 | 855.298584 | 10440.669241 | 4357.526384 | 549.380952 | 1340.642214 | 7 |
| **std** | 3870.201484 | 2451.113971 | 929.176190 | 17.640364 | 19.804778 | 4850.420531 | 1522.431887 | 4023.016484 | 1096.696416 | 165.105360 | 677.071454 | 1 |
| **min** | 81.000000 | 72.000000 | 35.000000 | 1.000000 | 9.000000 | 139.000000 | 1.000000 | 2340.000000 | 1780.000000 | 96.000000 | 250.000000 | |
| **25%** | 776.000000 | 604.000000 | 242.000000 | 15.000000 | 41.000000 | 992.000000 | 95.000000 | 7320.000000 | 3597.000000 | 470.000000 | 850.000000 | 6 |
| **50%** | 1558.000000 | 1110.000000 | 434.000000 | 23.000000 | 54.000000 | 1707.000000 | 353.000000 | 9990.000000 | 4200.000000 | 500.000000 | 1200.000000 | 7 |
| **75%** | 3624.000000 | 2424.000000 | 902.000000 | 35.000000 | 69.000000 | 4005.000000 | 967.000000 | 12925.000000 | 5050.000000 | 600.000000 | 1700.000000 | 8 |
| **max** | 48094.000000 | 26330.000000 | 6392.000000 | 96.000000 | 100.000000 | 31643.000000 | 21836.000000 | 21700.000000 | 8124.000000 | 2340.000000 | 6800.000000 | 10 |




## Inference:

- No columns have Null values
- The average number of Applications is 3001.
- The number of Applications ranges from 81 to 48094.
- The average number of Acceptances is 2018.
- The number of Acceptances ranges from 72 to 26330.
- The average number of Faculties with PhD is 72.6.
- Applications have the highest correlation with Accept, Enroll and F.Undergrad.

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

### 2.2 Answer:

Scaling **is necessary** in this case as the values vary a lot by scale in different columns. Certain values are in the thousands and certain values are two digit only. Scaling is useful to compare these values. As PCA calculates a new projection based on standard deviation, values with high standard deviation will have a higher weight in the calculation.

The primary objective of scaling is to normalize the data in the given range. Scaling also helps speed up the algorithm.

| | 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 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **772** | -0.207906 | -0.205541 | -0.255036 | -1.335492 | -1.504518 | -0.125949 | 0.770939 | -0.905706 | -0.417186 | -0.299088 | -0.207721 | -0.775362 | -1.338284 | 1.745 |
| **773** | -0.269402 | -0.087227 | -0.091450 | -0.201728 | -0.444168 | -0.175430 | 0.165329 | 0.268289 | 0.549353 | 0.306586 | -0.133874 | 0.020809 | -0.319426 | -0.199 |
| **774** | -0.233745 | -0.042350 | -0.091450 | 0.365154 | 0.262732 | -0.186975 | -0.452762 | -0.880103 | -0.143637 | 0.409551 | -0.826563 | -0.346655 | -0.319426 | 0.078 |
| **775** | 1.990429 | 0.177142 | 0.577960 | 3.823132 | 2.181461 | 0.312775 | -0.507280 | 2.336389 | 1.962689 | 0.488288 | 1.143687 | 1.429419 | 1.106976 | -2.094 |
| **776** | -0.003266 | -0.066829 | -0.095755 | 0.025025 | 0.363718 | -0.146772 | 0.571915 | -1.354871 | -0.727208 | -0.299088 | -0.133874 | 0.143297 | -0.319426 | 1.013 |

777 rows × 17 columns

## 2.3 Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].

### 2.3 Answer:

- Correlation: measures both the strength and direction of the linear relationship between two variables.
- Covariance: is used to measure how variables change with respect to each other, indicating direction of relation.

```
In [175]: # Covariance Matrix:
          cov_mat = np.cov(df_3_std.T)
          cov_mat

Out[175]: array([[ 1.        ,  0.94345057,  0.84682205,  0.33883368,  0.3516399 ,
                   0.81449058,  0.39826427,  0.05015903,  0.16493896,  0.1325586 ,
                   0.17873085,  0.39069733,  0.36949147,  0.09563303, -0.09022589,
                   0.25959198,  0.1467546 ],
                 [ 0.94345057,  1.        ,  0.91163666,  0.19244693,  0.24747574,
                   0.87422328,  0.44127073, -0.02575455,  0.09089863,  0.11352535,
                   0.20098867,  0.35575788,  0.33758337,  0.17622901, -0.15998987,
                   0.12471701,  0.06731255],
                 [ 0.84682205,  0.91163666,  1.        ,  0.18129353,  0.22674511,
                   0.96463965,  0.5130686 , -0.15547734, -0.04023168,  0.11271089,
                   0.28092946,  0.33146914,  0.30827407,  0.23727131, -0.18079413,
                   0.06416923, -0.02234104],
                 [ 0.33883368,  0.19244693,  0.18129353,  1.        ,  0.89199497,
                   0.14128873, -0.10535628,  0.56233054,  0.37148038,  0.11885843,
                  -0.0933164 ,  0.53182802,  0.49113502, -0.38487451,  0.45548526,
                   0.66091341,  0.49498923],
                 [ 0.3516399 ,  0.24747574,  0.22674511,  0.89199497,  1.        ,
                   0.19944466, -0.05357664,  0.48939383,  0.33148989,  0.11552713,
                  -0.08081027,  0.54586221,  0.52474884, -0.29462884,  0.41786429,
                   0.52744743,  0.47728116],
                 [ 0.81449058,  0.87422328,  0.96463965,  0.14128873,  0.19944466,
                   1.        ,  0.57051219, -0.215742  , -0.06889039,  0.11554976,
                   0.31719954,  0.31833697,  0.30001894,  0.27970335, -0.22946222,
                   0.01865162, -0.07877313],
                 [ 0.39826427,  0.44127073,  0.5130686 , -0.10535628, -0.05357664,
                   0.57051219,  1.        , -0.25351232, -0.06132551,  0.08119952,
                   0.31988162.  0.14911422.  0.14190357.  0.23253051. -0.28079236.
```

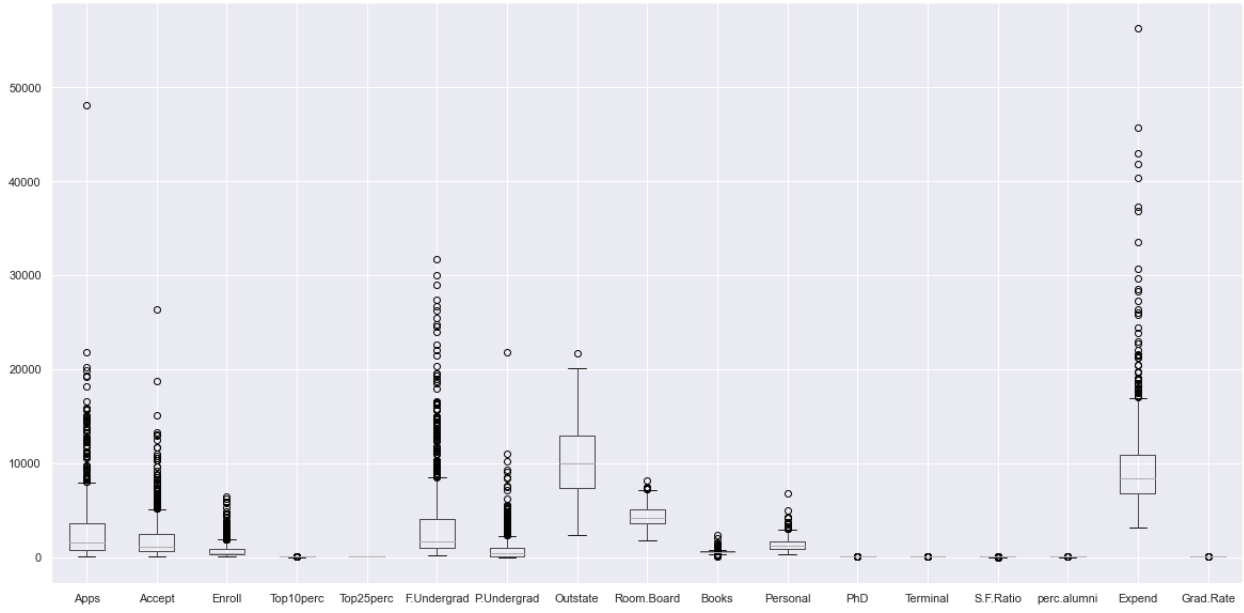
```
In [165]: # Correlation Matrix

          df_3_std.corr()

Out[165]:
```

| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal | PhD | Tern |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **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 | 0.36 |
| **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 | 0.33 |
| **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 | 0.30 |
| **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 | 0.49 |
| **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 | 0.52 |
| **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 | 0.30 |
| **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 | 0.14 |
| **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 | 0.40 |
| **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 | 0.37 |
| **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 | 0.09 |
| **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 | -0.03 |

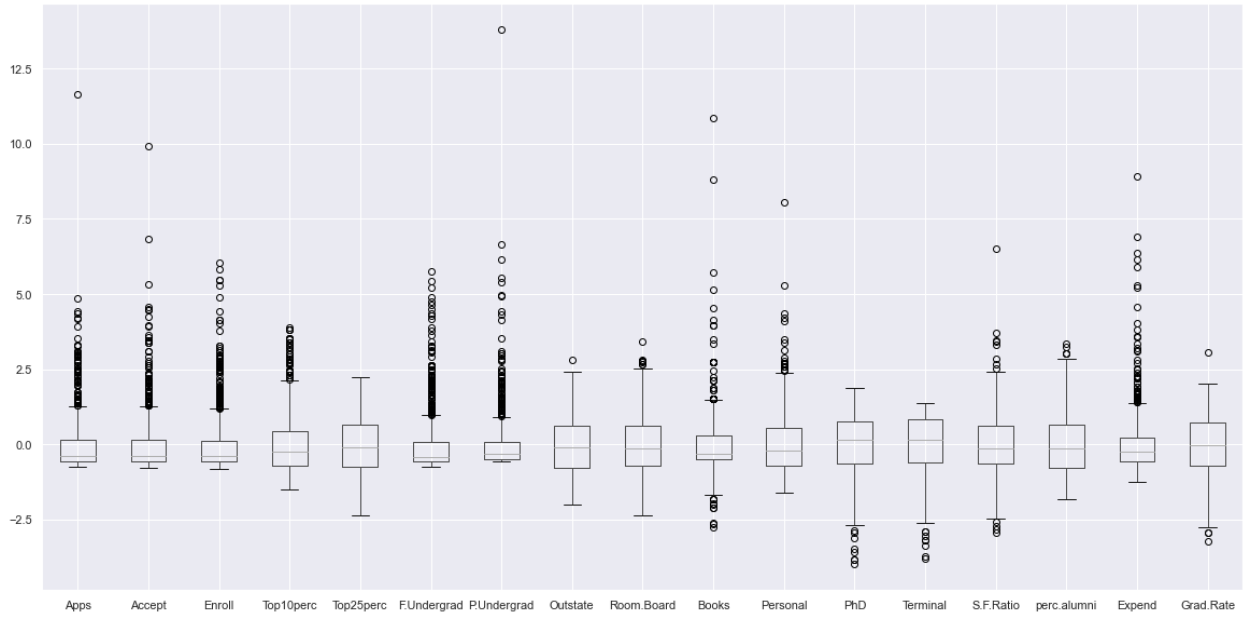
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

**2.4 Answer:**

**Before Scaling:**




## After Scaling:




**Inference:** Scaling reduces the number of outliers

2.5 Extract the eigenvalues and eigenvectors.[print both]

**2.5 Answer:**

```
Eigen Values are:


array([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])


 Eigen Vectors are:


array([[-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
```

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

**2.6 Answer:**

```
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]])
```

2.7 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]

**2.7 Answer:**

[ -0.25, 0.33, 0.06,

-0.28, 0.01, 0.02,

0.04, 0.10, 0.09,

-0.05, 0.36, -0.46,

0.04, -0.13, 0.08,

-0.60, 0.02 ]

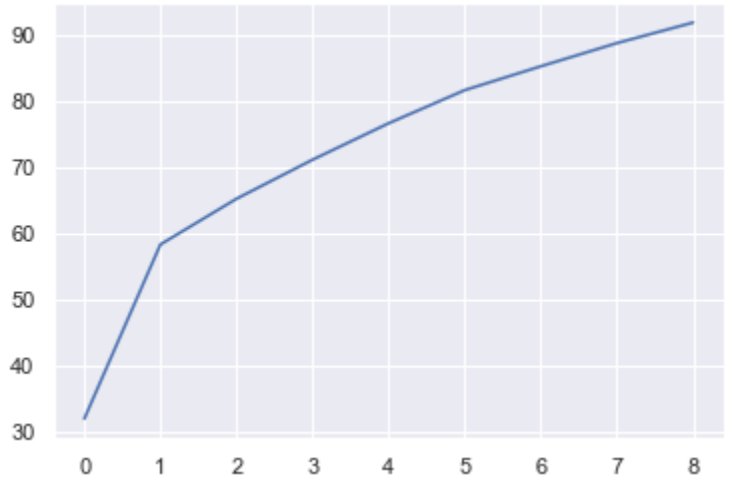
2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

**2.8 Answer:**

Cumulative Percentage helps us decide how many components to select by showing the percentage of variance that those components were able to show.

It helps us decide the optimum number of principal components by making sure the correct variance in data is represented.

For example: In the diagram below having 9 principal components helps us show more than 90% of the variance, hence 9 may be the ideal number of principal components.



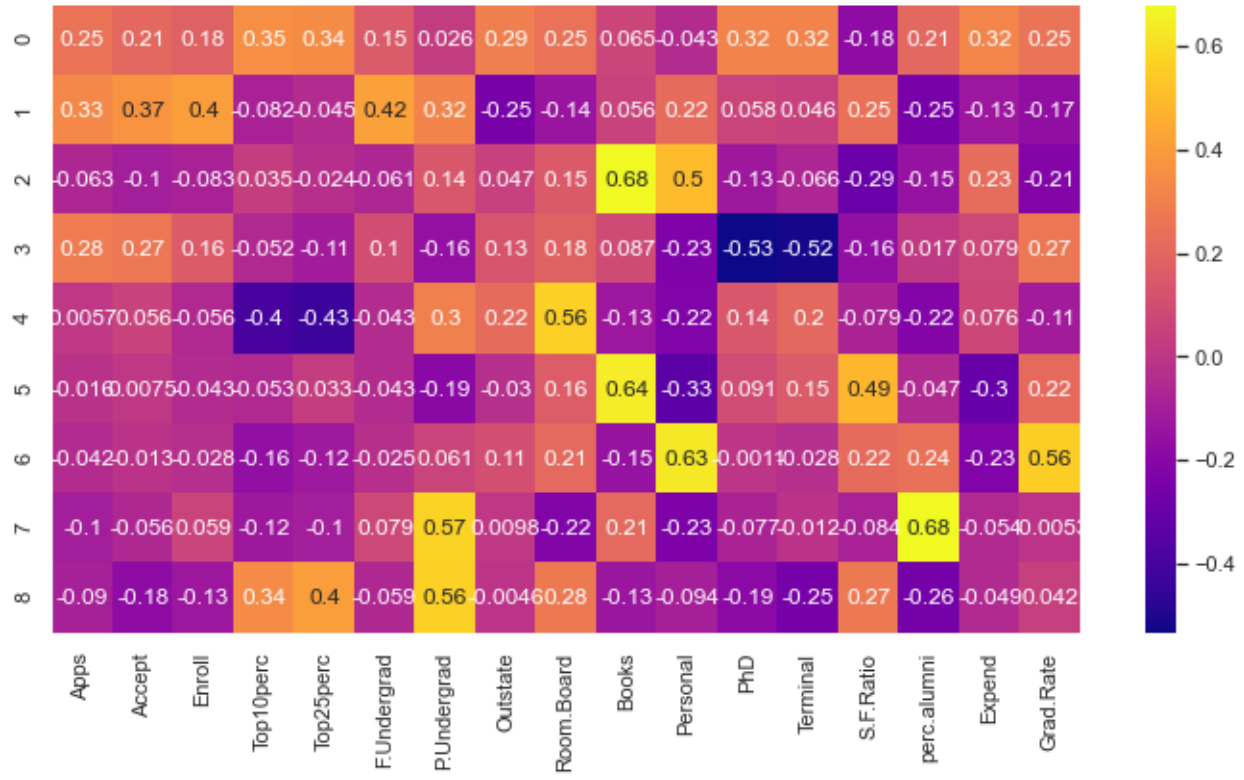

[32. , 58.3, 65.2, 71.1, 76.6, 81.6, 85.2, 88.7, 91.8]

**Eigen Vectors** indicate the direction of the Principal Components. We can multiply the original data by the them to re-orient our data onto the new axes.

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]

**2.9 Answer:**

- 9 Principal components are enough to represent more than 90% of the variance in the data.
- The dimensionality of the data can be reduced from 17 to 9.
- PCA 0 explains: Explains Top10perc and Top25perc.
- PCA 1 explains: Enroll and F.Undergrad.




- PCA 2 explains: Books and Personal.
- PCA 3 explains: PhD and Terminal.
- PCA 4 explains: Top10perc, Top25perc and Room.Board.
- PCA 5 explains: Books and Personal.
- PCA 6 explains: Personal and Grad Rate.
- PCA 7 explains: perc.alumini and P.Undergrad.
- PCA 8 explains: P.Undergrad and Top25perc.