Problem 2

The dataset Education - Post 12th Standard.csv is a dataset that contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. 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]. The inferences drawn from this should be properly documented.

I have done some descriptive analysis to check the data types, size and to check if there are any missing or duplicated values by using `isna()` function and `duplicated()` function. There are no missing or duplicated values.

For doing univariate analysis, I have used `distplot` and `boxplot` functions in `seaborn` to know the distribution and outliers. I have performed univariate analysis on all the columns except the `names` column which is an object datatype.

The inferences drawn from the plots is that all the variables have outliers except `Top25perc`. The variables `Apps`, `Accept`, `Enroll`, `F.Undergrad`, `P.Undergrad`, `expend` are positively distributed and the variables `Top 10 per`, `Room board`, `Personal` are slightly positively distributed. The variables `Top 25 per`, `Grad rate` shows normal distribution but grad rate has outliers. The variables `Outstate`, `S.F.ratio`, per alumni show slightly normal distribution. The variables `PhD`, `Terminal` is negatively skewed and the variable books shows uneven distribution with a lot of outliers.

For multivariate analysis, I have checked the correlation by using the `Pearson` method and plotted the heat map in the `seaborn` package. `S.F.ratio` and `Expend` are poorly correlated with a correlation coefficient of `-0.58` which means there is no dependency between Student/faculty ratio and the expenditure per student. Enroll and `f` undergrad are highly correlated with a correlation coefficient of `0.96` showing that in the number of students enrolled, most of them are full time undergraduate students. And also, most of the applications are accepted, and the enrolled students contain more `top10` and `top25` higher secondary school graduates. And the people coming from outstate may prefer room boards since they are somehow correlated. And also, most of the applications received are for full time undergraduate students.

I have checked the distribution and relation using a `pair plot` in `seaborn` package. And I have used boxplot to check for outliers. Except `Top25 perc`, all columns have outliers.

I have defined a custom function to treat outliers in which if for a particular column the maximum value is greater than that assigned maximum value, minimum value is lower than that assigned minimum value.

2.2) Scale the variables and write the inference for using the type of scaling function for this case study.

Scaling or standardization is done to make all the variables on the same scale. By standardizing, we find the zero mean of each column, by subtracting the mean from each row for every column in our dataset. Next, we divide through by the standard deviation to have a specific range of numbers.

Here I dropped the name variable which is of `object` data type by using `drop` function and standardize the data using `zscore` from `SciPy` package.

By plotting the new standardized data using boxplot from `seaborn`, we observed that there are outliers in the columns `Top10perc`, `S.F.Ratio`, `perc. Alumni`. Almost all columns are positively skewed except `top25perc`, `outstate`, `grad rate` which seems to be symmetric and terminal, `personal` seems to be negatively skewed.

2.3) Comment on the comparison between covariance and the correlation matrix.

Both Correlation and Covariance are very closely related to each other and yet they differ a lot.

When it comes to choosing between Covariance vs Correlation, correlation stands to be the first choice as it remains unaffected by the change in dimensions, location, and scale, and can also be used to make a comparison between two pairs of variables. With standardization (Without standardization also, correlation matrix yields the same result. Since it is limited to a range of -1 to +1, it is useful to draw comparisons between variables across domains. “Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. You can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.

But the limitation is that both these concepts measure only linear relationships.

After scaling, the covariance and the correlation have the same values.

2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

I used the boxplot function in seaborn to check for outliers. The inferences drawn from the plots is that all the variables have outliers except `Top25perc`. I have defined a custom function to treat outliers in which if for a particular column the maximum value is greater than that assigned maximum value, minimum value is lower than that assigned minimum value. After scaling, by plotting the new standardized data using `boxplot` from `seaborn`, we observed that there are outliers in the columns `Top10perc`, `S.F.Ratio`, `perc. Alumni`.

2.5) Build the covariance matrix, eigenvalues, and eigenvector.

The covariance matrix of the dataset is calculated by multiplying the matrix of features by its transpose. It is a measure of how much each of the dimensions vary from the mean with respect to each other.

A positive value of covariance indicates that both the dimensions are directly proportional to each other, where if one dimension increases the other dimension increases accordingly.

A negative value of covariance indicates that both the dimensions are indirectly proportional to each other, where if one dimension increases then the other dimension decreases accordingly.

If in case the covariance is zero, then the two dimensions are independent of each other.

In the covariance matrix in the output, the off-diagonal elements contain the covariances of each pair of variables. The diagonal elements of the covariance matrix contain the variances of each variable. The variance measures how much the data are scattered about the mean.

```
Covariance Matrix [[ 1.00128866e+00 9.56537704e-01 8.98039052e-01 3.21756324e-01
3.64960691e-01 8.62111140e-01 5.20492952e-01 6.54209711e-02   1.87717056e-01
2.36441941e-01 2.30243993e-01 4.64521757e-01   4.35037784e-01 1.26573895e-01
-1.01288006e-01  2.43248206e-01   1.50997775e-01] [ 9.56537704e-01  1.00128866e+00
9.36482483e-01  2.23586208e-01   2.74033187e-01  8.98189799e-01  5.73428908e-01
-5.00874847e-03   1.19740419e-01  2.08974091e-01  2.56676290e-01  4.27891234e-01
4.03929238e-01  1.88748711e-01 -1.65728801e-01  1.62016688e-01   7.90839722e-02] [
8.98039052e-01  9.36482483e-01  1.00128866e+00  1.71977357e-01   2.30730728e-01
9.68548601e-01  6.42421828e-01 -1.55856056e-01  -2.38762560e-02  2.02317274e-01
3.39785395e-01  3.82031198e-01   3.54835877e-01  2.74622251e-01 -2.23009677e-01
5.42906862e-02  -2.32810071e-02] [ 3.21756324e-01  2.23586208e-01  1.71977357e-01
1.00128866e+00   9.15052977e-01  1.11358019e-01 -1.80240778e-01  5.62884044e-01
3.57826139e-01  1.53650150e-01 -1.16880152e-01  5.44748764e-01   5.07401238e-01
-3.88425719e-01  4.56384036e-01  6.57885921e-01   4.94306540e-01] [ 3.64960691e-01
2.74033187e-01  2.30730728e-01  9.15052977e-01   1.00128866e+00  1.81429267e-01
-9.94231153e-02  4.90200034e-01   3.31413314e-01  1.69979808e-01 -8.69219644e-02
5.52172085e-01   5.28333659e-01 -2.97616423e-01  4.17369123e-01  5.73643193e-01
4.79601950e-01] [ 8.62111140e-01  8.98189799e-01  9.68548601e-01  1.11358019e-01
1.81429267e-01  1.00128866e+00  6.97027420e-01 -2.26457040e-01  -5.45459528e-02
2.08147257e-01  3.60246460e-01  3.62030390e-01   3.35485771e-01  3.24921933e-01
-2.85825062e-01  3.71119607e-04  -8.23447851e-02] [ 5.20492952e-01  5.73428908e-01
6.42421828e-01 -1.80240778e-01  -9.94231153e-02  6.97027420e-01  1.00128866e+00
-3.54672874e-01  -6.77252009e-02  1.22686416e-01  3.44495974e-01  1.27827147e-01
1.22309141e-01  3.71084841e-01 -4.19874031e-01 -2.02189396e-01  -2.65499420e-01]
[ 6.54209711e-02 -5.00874847e-03 -1.55856056e-01  5.62884044e-01   4.90200034e-01
-2.26457040e-01 -3.54672874e-01  1.00128866e+00   6.56333564e-01  5.11656377e-03
-3.26028927e-01  3.91824814e-01   4.13110264e-01 -5.74421963e-01  5.66465309e-01
7.76326650e-01   5.73195743e-01] [ 1.87717056e-01  1.19740419e-01 -2.38762560e-02
3.57826139e-01   3.31413314e-01 -5.45459528e-02 -6.77252009e-02  6.56333564e-01
```

```
1.00128866e+00  1.09064551e-01 -2.19837042e-01  3.41908577e-01   3.79759015e-01
-3.76915472e-01  2.72743761e-01  5.81370284e-01   4.26338910e-01] [ 2.36441941e-01
2.08974091e-01  2.02317274e-01  1.53650150e-01   1.69979808e-01  2.08147257e-01
1.22686416e-01  5.11656377e-03   1.09064551e-01  1.00128866e+00  2.40172145e-01
1.36566243e-01   1.59523091e-01 -8.54689129e-03 -4.28870629e-02  1.50176551e-01
-8.06107505e-03] [ 2.30243993e-01  2.56676290e-01  3.39785395e-01 -1.16880152e-01
-8.69219644e-02  3.60246460e-01  3.44495974e-01 -3.26028927e-01  -2.19837042e-01
2.40172145e-01  1.00128866e+00 -1.16986124e-02  -3.20117803e-02  1.74136664e-01
-3.06146886e-01 -1.63481407e-01  -2.91268705e-01] [ 4.64521757e-01  4.27891234e-01
3.82031198e-01  5.44748764e-01   5.52172085e-01  3.62030390e-01  1.27827147e-01
3.91824814e-01   3.41908577e-01  1.36566243e-01 -1.16986124e-02  1.00128866e+00
8.64040263e-01 -1.29556494e-01  2.49197779e-01  5.11186852e-01   3.10418895e-01] [
4.35037784e-01  4.03929238e-01  3.54835877e-01  5.07401238e-01   5.28333659e-01
3.35485771e-01  1.22309141e-01  4.13110264e-01   3.79759015e-01  1.59523091e-01
-3.20117803e-02  8.64040263e-01   1.00128866e+00 -1.51187934e-01  2.66375402e-01
5.24743500e-01   2.93180212e-01] [ 1.26573895e-01  1.88748711e-01  2.74622251e-01
-3.88425719e-01  -2.97616423e-01  3.24921933e-01  3.71084841e-01 -5.74421963e-01
-3.76915472e-01 -8.54689129e-03  1.74136664e-01 -1.29556494e-01  -1.51187934e-01
1.00128866e+00 -4.12632056e-01 -6.55219504e-01  -3.08922187e-01] [-1.01288006e-01
-1.65728801e-01 -2.23009677e-01  4.56384036e-01   4.17369123e-01 -2.85825062e-01
-4.19874031e-01  5.66465309e-01   2.72743761e-01 -4.28870629e-02 -3.06146886e-01
2.49197779e-01   2.66375402e-01 -4.12632056e-01  1.00128866e+00  4.63518674e-01
4.92040760e-01] [ 2.43248206e-01  1.62016688e-01  5.42906862e-02  6.57885921e-01
5.73643193e-01  3.71119607e-04 -2.02189396e-01  7.76326650e-01   5.81370284e-01
1.50176551e-01 -1.63481407e-01  5.11186852e-01   5.24743500e-01 -6.55219504e-01
4.63518674e-01  1.00128866e+00   4.15826026e-01] [ 1.50997775e-01  7.90839722e-02
-2.32810071e-02  4.94306540e-01   4.79601950e-01 -8.23447851e-02 -2.65499420e-01
5.73195743e-01   4.26338910e-01 -8.06107505e-03 -2.91268705e-01  3.10418895e-01
2.93180212e-01 -3.08922187e-01  4.92040760e-01  4.15826026e-01   1.00128866e+00]]

Eigenvalues and eigenvectors allow us to reduce a linear operation to separate,
simpler, problems. They can be obtained by using #eig_vals, eig_vecs =
np.linalg.eig(cov_matrix)

Eigen Values  %s [5.6625219  4.89470815 1.12636744 1.00397659 0.87218426 0.7657541
0.58491404 0.5445048  0.42352336 0.38101777 0.24701456 0.02239369 0.03789395
0.14726392 0.13434483 0.09883384 0.07469003]

Eigen Vectors  %s [[-2.62171542e-01  3.14136258e-01  8.10177245e-02 -9.87761685e-02
-2.19898081e-01  2.18800617e-03 -2.83715076e-02 -8.99498102e-02   1.30566998e-01
-1.56464458e-01 -8.62132843e-02  1.82169814e-01  -5.99137640e-01  8.99775288e-02
8.88697944e-02  5.49428396e-01   5.41453698e-03] [-2.30562461e-01  3.44623583e-01
1.07658626e-01 -1.18140437e-01  -1.89634940e-01 -1.65212882e-02 -1.29584896e-02
-1.37606312e-01   1.42275847e-01 -1.49209799e-01 -4.25899061e-02 -3.91041719e-01
6.61496927e-01  1.58861886e-01  4.37945938e-02  2.91572312e-01   1.44582845e-02]
[-1.89276397e-01  3.82813322e-01  8.55296892e-02 -9.30717094e-03  -1.62314818e-01
-6.80794143e-02 -1.52403625e-02 -1.44216938e-01   5.08712481e-02 -6.48997860e-02
```

```
-4.38408622e-02  7.16684935e-01   2.33235272e-01 -3.53988202e-02 -6.19241658e-02
-4.17001280e-01  -4.97908902e-02] [-3.38874521e-01 -9.93191661e-02 -7.88293849e-02
3.69115031e-01  -1.57211016e-01 -8.88656824e-02 -2.57455284e-01  2.89538833e-01
-1.22467790e-01 -3.58776186e-02  1.77837341e-03 -5.62053913e-02   2.21448729e-02
-3.92277722e-02  6.99599977e-02  8.79767299e-03  -7.23645373e-01] [-3.34690532e-01
-5.95055011e-02 -5.07938247e-02  4.16824361e-01  -1.44449474e-01 -2.76268979e-02
-2.39038849e-01  3.45643551e-01  -1.93936316e-01  6.41786425e-03 -1.02127328e-01
1.96735274e-02   3.22646978e-02  1.45621999e-01 -9.70282598e-02 -1.07779150e-02
6.55464648e-01] [-1.63293010e-01  3.98636372e-01  7.37077827e-02 -1.39504424e-02
-1.02728468e-01 -5.16468727e-02 -3.11751439e-02 -1.08748900e-01   1.45452749e-03
-1.63981359e-04 -3.49993487e-02 -5.42774834e-01  -3.67681187e-01 -1.33555923e-01
-8.71753137e-02 -5.70683843e-01   2.53059904e-02] [-2.24797091e-02  3.57550046e-01
4.03568700e-02 -2.25351078e-01   9.56790178e-02 -2.45375721e-02 -1.00138971e-02
1.23841696e-01  -6.34774326e-01  5.46346279e-01  2.52107094e-01  2.95029745e-02
2.62494456e-02  5.02487566e-02  4.45537493e-02  1.46321060e-01  -3.97146972e-02]
[-2.83547285e-01 -2.51863617e-01  1.49394795e-02 -2.62975384e-01  -3.72750885e-02
-2.03860462e-02  9.45370782e-02  1.12721477e-02  -8.36648339e-03 -2.31799759e-01
5.93433149e-01  1.03393587e-03  -8.14247697e-02  5.60392799e-01  6.72405494e-02
-2.11561014e-01  -1.59275617e-03] [-2.44186588e-01 -1.31909124e-01 -2.11379165e-02
-5.80894132e-01   6.91080879e-02  2.37267409e-01  9.45210745e-02  3.89639465e-01
-2.20526518e-01 -2.55107620e-01 -4.75297296e-01  9.85725168e-03   2.67779296e-02
-1.07365653e-01  1.77715010e-02 -1.00935084e-01  -2.82578388e-02] [-9.67082754e-02
9.39739472e-02 -6.97121128e-01  3.61562884e-02  -3.54056654e-02  6.38604997e-01
-1.11193334e-01 -2.39817267e-01   2.10246624e-02  9.11624912e-02  4.35697999e-02
4.36086500e-03   1.04624246e-02  5.16224550e-02  3.54343707e-02 -2.86384228e-02
-8.06259380e-03] [ 3.52299594e-02  2.32439594e-01 -5.30972806e-01  1.14982973e-01
4.75358244e-04 -3.81495854e-01  6.39418106e-01  2.77206569e-01   1.73715184e-02
-1.27647512e-01  1.51627393e-02 -1.08725257e-02   4.54572099e-03  9.39409228e-03
-1.18604404e-02  3.38197909e-02   1.42590097e-03] [-3.26410696e-01  5.51390195e-02
8.11134044e-02  1.47260891e-01   5.50786546e-01  3.34444832e-03  8.92320786e-02
-3.42628480e-02   1.66510079e-01  1.00975002e-01 -3.91865961e-02  1.33146759e-02
1.25137966e-02 -7.16590441e-02  7.02656469e-01 -6.38096394e-02   8.31471932e-02]
[-3.23115980e-01  4.30332048e-02  5.89785929e-02  8.90079921e-02   5.90407136e-01
3.54121294e-02  9.16985445e-02 -9.03076644e-02   1.12609034e-01  8.60363025e-02
-8.48575651e-02  7.38135022e-03  -1.79275275e-02  1.63820871e-01 -6.62488717e-01
9.85019644e-02  -1.13374007e-01] [ 1.63151642e-01  2.59804556e-01  2.74150657e-01
2.59486122e-01   1.42842546e-01  4.68752604e-01  1.52864837e-01  2.42807562e-01
-1.53685343e-01 -4.70527925e-01  3.63042716e-01  8.85797314e-03   1.83059753e-02
-2.39902591e-01 -4.79006197e-02  6.19970446e-02   3.83160891e-03] [-1.86610828e-01
-2.57092552e-01  1.03715887e-01  2.23982467e-01  -1.28215768e-01  1.25669415e-02
3.91400512e-01 -5.66073056e-01  -5.39235753e-01 -1.47628917e-01 -1.73918533e-01
-2.40534190e-02  -8.03169296e-05 -4.89753356e-02  3.58875507e-02  2.80805469e-02
-7.32598621e-03] [-3.28955847e-01 -1.60008951e-01 -1.84205687e-01 -2.13756140e-01
2.24240837e-02 -2.31562325e-01 -1.50501305e-01 -1.18823549e-01   2.42371616e-02
-8.04154875e-02  3.93722676e-01  1.05658769e-02   5.60069250e-02 -6.90417042e-01
-1.26667522e-01  1.28739213e-01   1.45099786e-01] [-2.38822447e-01 -1.67523664e-01
2.45335837e-01  3.61915064e-02  -3.56843227e-01  3.13556243e-01  4.68641965e-01
```

```
1.80458508e-01   3.15812873e-01  4.88415259e-01  8.72638706e-02 -2.51028410e-03
1.48410810e-02 -1.59332164e-01 -6.30737002e-02 -7.09643331e-03  -3.29024228e-03]]
```

2.6) Write the explicit form of the first PC (in terms of Eigen Vectors).

Line of best fit drawn representing the direction of the first eigenvector, which is the first PCA component.

The main principal component, depicted by the black line, is the first Eigenvector. The second Eigenvector will be perpendicular or orthogonal to the first one. The reason the two Eigenvectors are orthogonal to each other is because the Eigenvectors should be able to span the whole x-y area. Naturally, a line perpendicular to the black line will be our new Y axis, the other principal component.

2.7) Discuss 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?

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

Cumulative distribution can be explained with the help of `scree` plot. Visually we can observe that there is a steep drop in cumulative variance with an increase in the number of PC's. Based on Scree Plot, we can decide on the optimum number of principal components. Principal components with eigenvalues greater than 1 are considered more significant. Eigenvectors indicate the direction of maximum variance in the dataset.

Here we are generating only 4 PCA dimensions. Using `scikit` learn PCA here, it does all the steps and maps data to PCA dimensions in one shot. We will take the transpose of the reduced data to get our new data frame. To select the components with high variance, we will find cumulative percentage which gives the percentage of sum of variance explained with `[n]` features.

```
array ([33.3, 62.1, 68.7, 74.6])
```

In the above output, we see that the first feature explains `33.3%` of the variance within our data set while the first two explain `62.1` and so on. If we employ `4` features, we capture approximately `74.6%` of the variance within the dataset, thus we gain very little by implementing an additional feature.

PCA can only be done on continuous variables

2.8) Mention the business implication of using the Principal Component Analysis for this case study. [Hint: Write Interpretations of the Principal Components Obtained]

PCA is a method that brings together measures of how each variable is associated with one another (covariance matrix), the directions in which our data are dispersed (Eigenvectors), the relative importance of these different directions (Eigenvalues). The idea is to re-align the axis in an n-dimensional space such that we can capture most of the variance in the data. Here we can see in the heatmap that the maximum variance is in the estimated books cost followed by estimated personal spending and cost of room and board per student.