**PROBLEM 1A:-**

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

Ans:- for variable Education:-

Ho = Mean salary of people with Doctorate = Mean salary of people with Bachelors = Mean Salary of people with HS-grad

Ha = Means salaries are not equal with respect to different education level

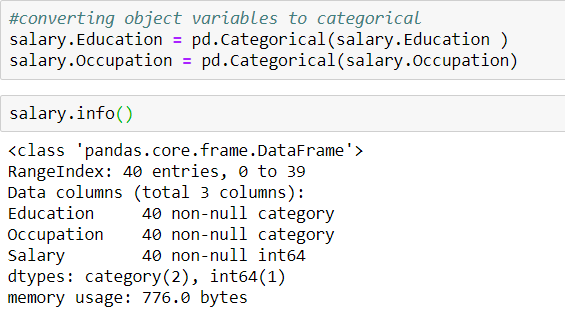
For Variable Occupation:-

Ho = Mean salary of Adm-clerical = Mean salary of Exec-managerial = Mean Salary of Prof-specialty = Mean Salary of Sales

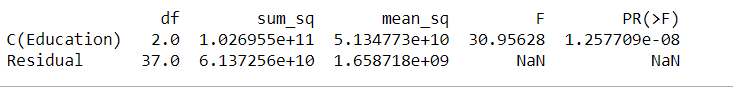
Ha = Means salaries are not equal with respect to different Occupation

**1.2 Perform one-way ANOVA for Education with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**

Ans:- We need to convert education and Occupation into categorical data. Below is the result:-



Below is the result of one-way ANOVA test:-



Based on one way ANOVA result above PR(>F) = 1.257709e-08. This indicate 'p' value is less than the significance level of 0.05.

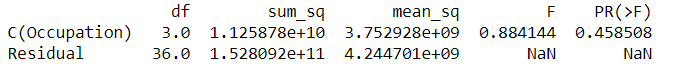
We can reject the null hypothesis and states that 'at least one of the three means' are not equal. We can conclude that the

different Education level have significant effect over Salary.

**1.3 Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**

Ans:-

We have already converted Occupation variable into categorical as shown in previous question. Below are the result of one-way ANOVA test:-

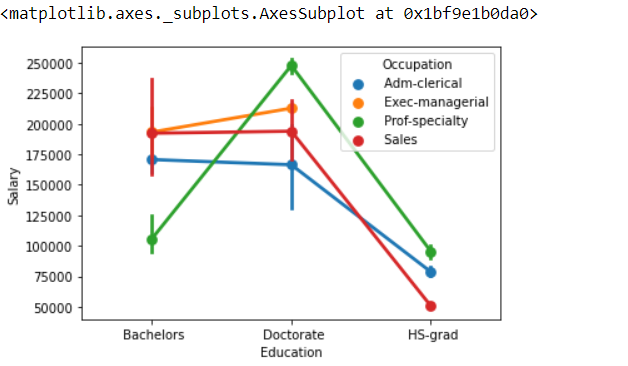


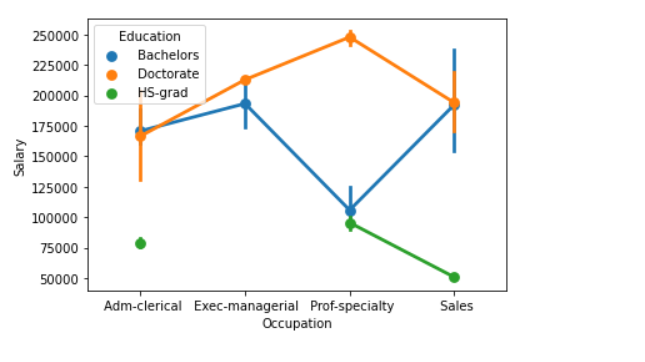
Based on one way ANOVA result above PR(>F) = 0.458508.this indicate 'p' value is more than the significance level of 0.05. We fail to reject the null hypothesis and states that mean of salaries across different Occupation is equal. We can conclude that the different Occupation does not have significant effect on Salary.

**PROBLEM 1B:-**

**1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.**

**Ans:-**



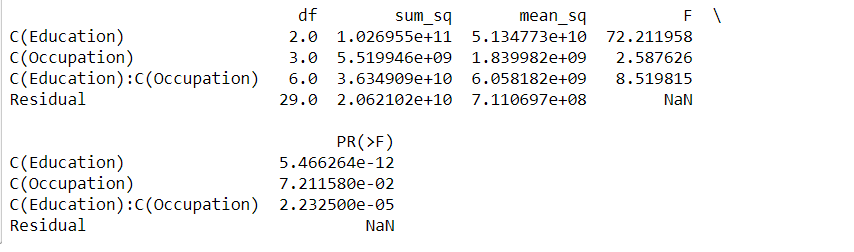


**1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education\*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result?**

**Ans:-**

**Ho = There is no interaction**

**Ha = There is interaction**



Since the P value is less than significance level of 0.05. we can reject the null hypothesis and conclude that the variable Education and Occupation seems to be statistically interacted.

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

Ans:-

* Considering individual effects of variable’s Education and Occupation we conclude that variable education has a significant effect on salary. However, Occupation doesn’t show the same.
* When considered together we found that Education\*Occupation have significant effect on Salary and that turns out to be true as individuals with higher education and occupation in the similar line of education cannot have same salary.

**PROBLEM 2:-**

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

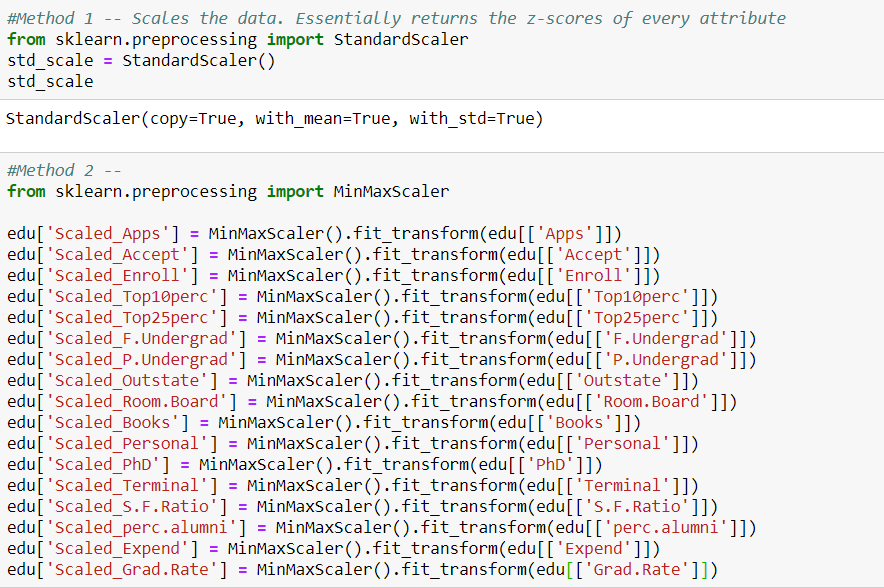
**Ans:-**

* We have 18 variables with 777 observations, Variable name is object datatype, S.F.Ratio variable is float, rest all variables are int64 data type.
* Checking for null values in data and found no null value in the data.
* Checked for duplicate date and found no duplication in the data.
* Visualized data via box plot and found that almost all the variables consist of outliers which needs treatment for further analysis.
* Almost all the variables are skewed in nature.
* Checked SD for variables to check how far they are form mean.
* Checked pair plot
* Checked Heat map
* Checked covariance and correlation matrix of the data.
  1. **Is scaling necessary for PCA in this case? Give justification and perform scaling.**

Ans:-

Yes, this case needs scaling as different variable have different units of measurement. So that we can normalize the range of all features.

We can perform scaling by below two methods:-

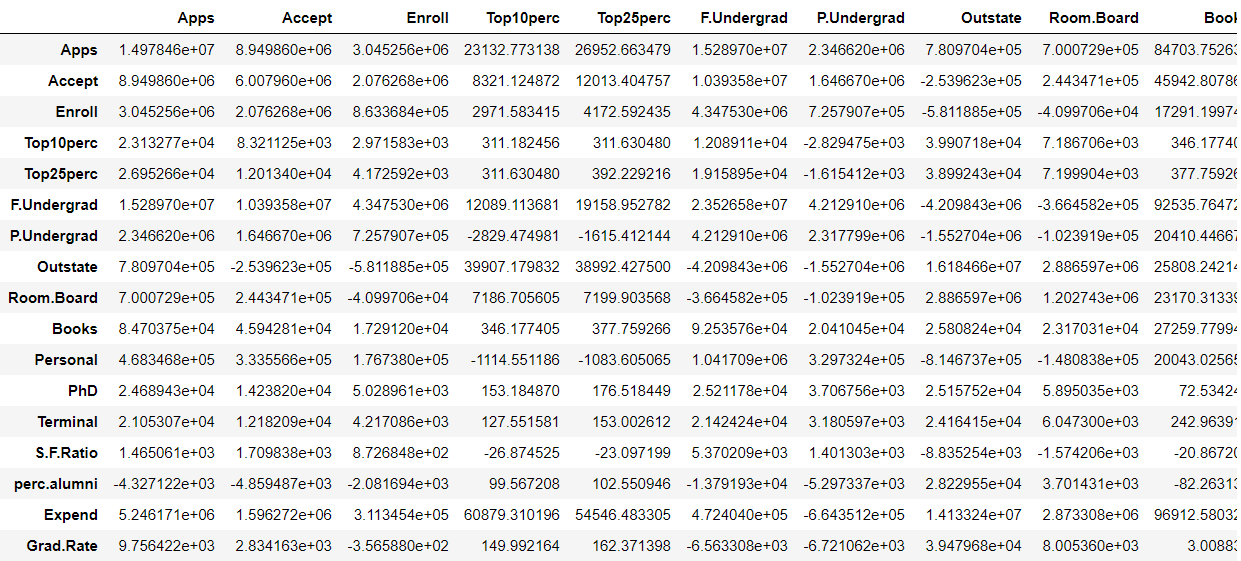


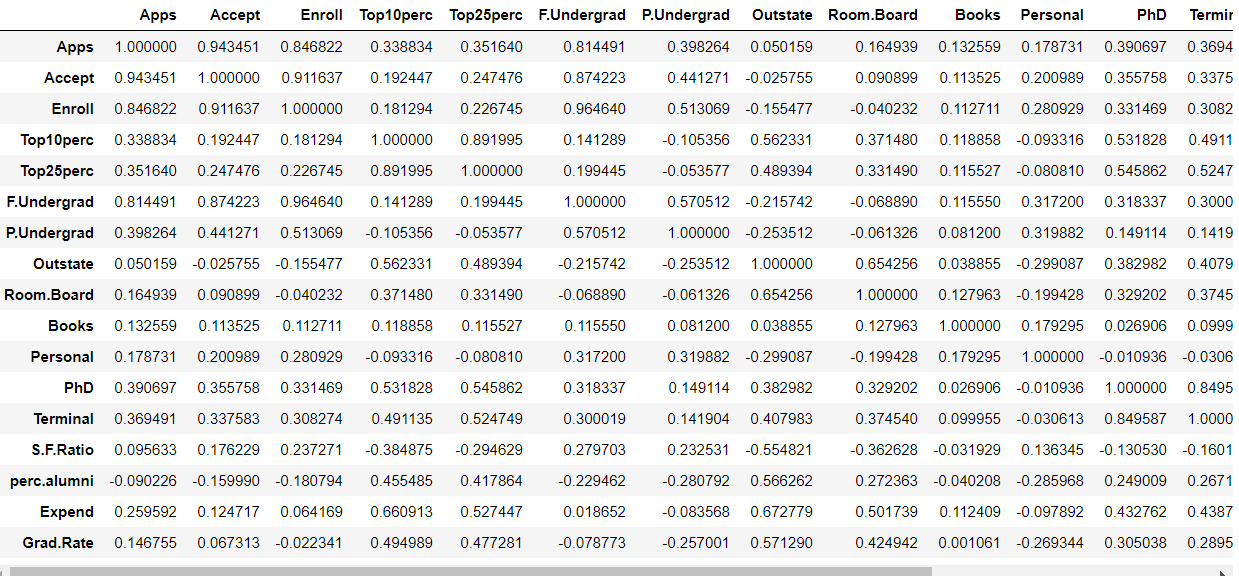
For this case I have MInMax scaler and created a different matrix for scales dataset.

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

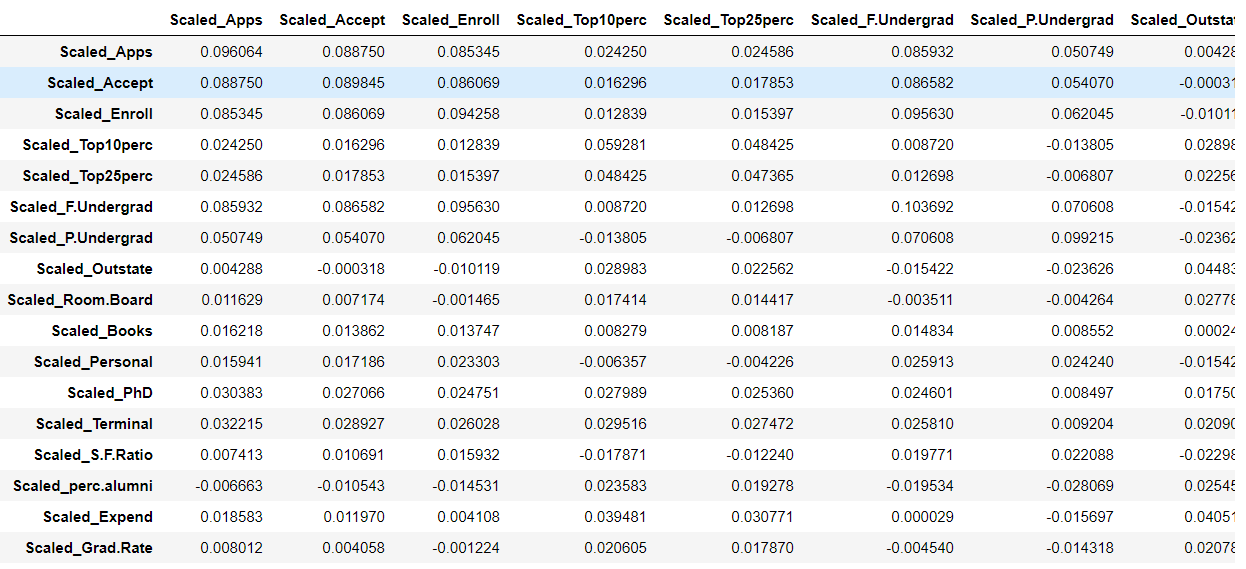
Ans:- Data is highly distributed and we can see both positive and negative covariance in the data. Effected by the same we see the distribution in correlation as well. However, there is change in the cov and corr post scaling which makes data look easy for analysis. Below are the details of Cov and Corr pre- and post-scaling.

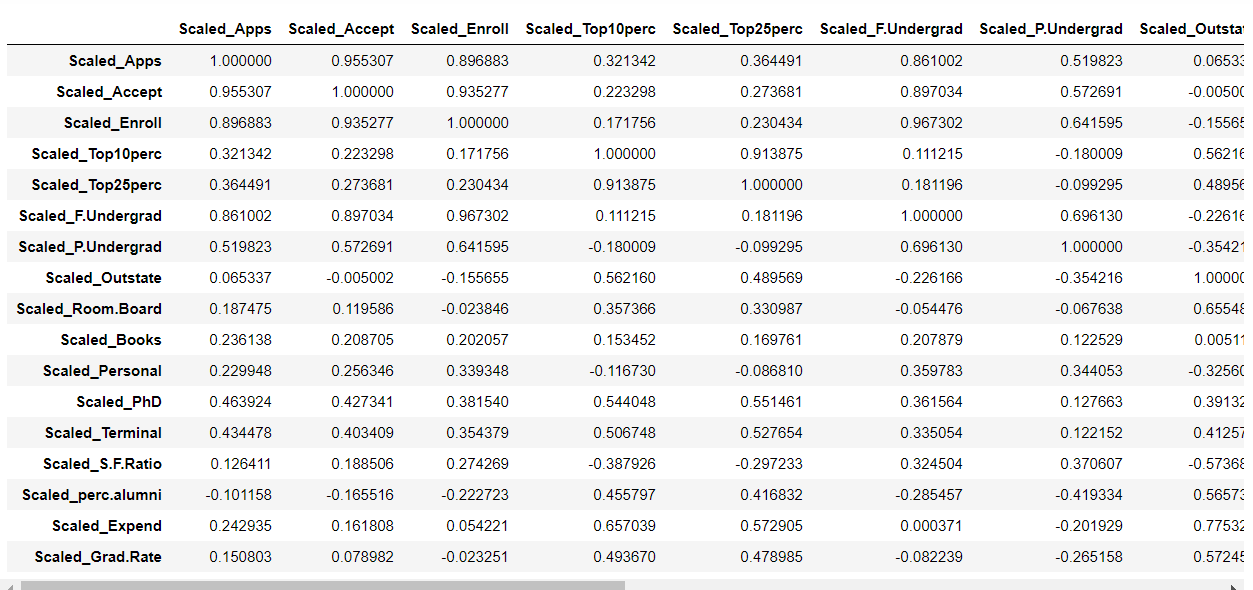
Cov and corr before scaling: -





Cov and corr after scaleing:-



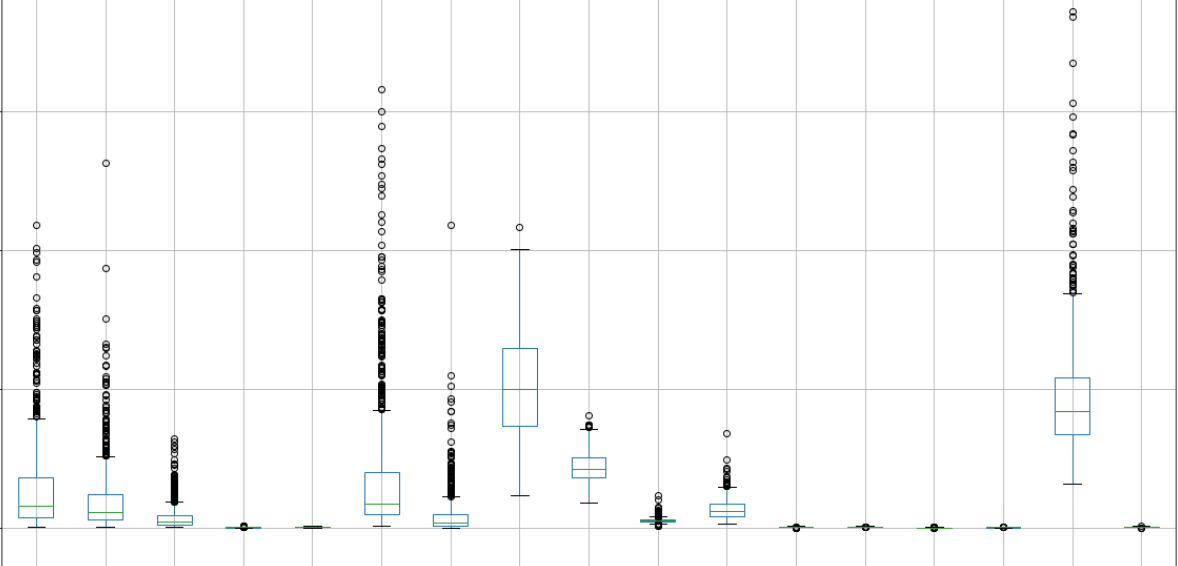


* 1. **Check the dataset for outliers before and after scaling. What insight do you derive here?**

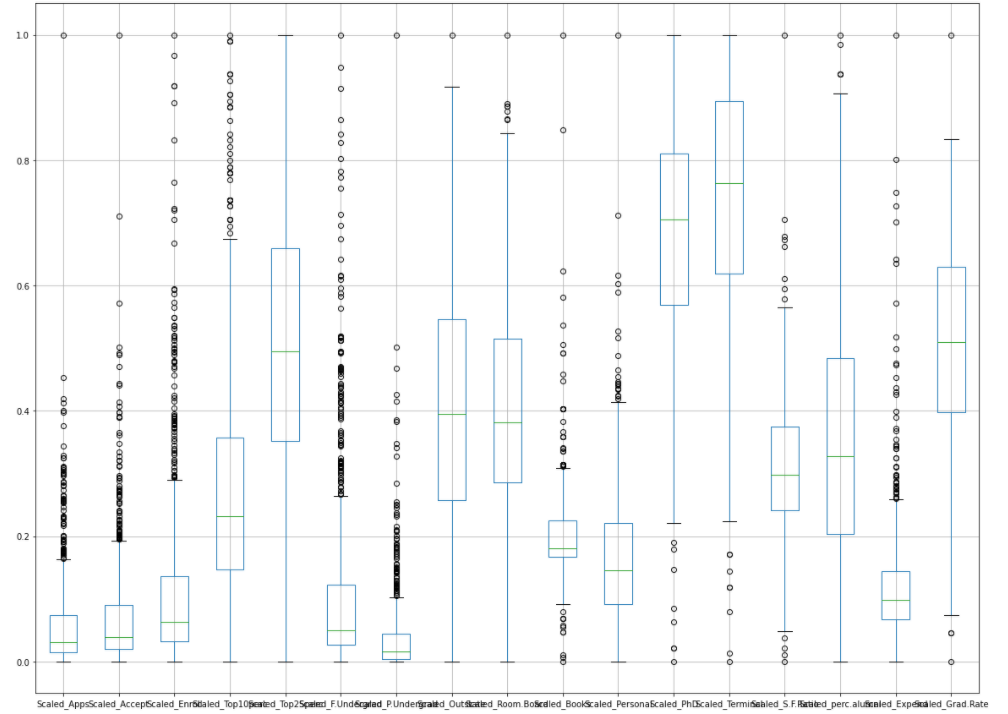
Ans:-

Below are the observations from the data for outliers post and pre scaling:-

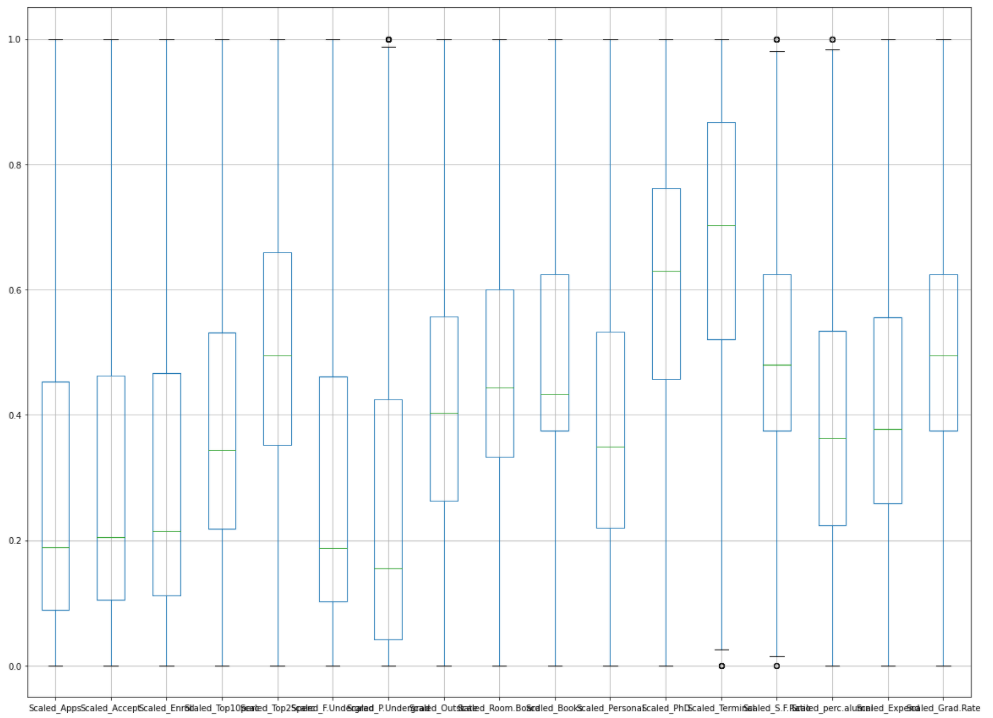
* Scaling helps data to be near to mean with no standard deviation and helps data to get into normal distribution.
* In the given case we still see skewness in the data because of the presence of outliers in the data.
* Post outlier treatment we can see that the data is normally distributed and ready for further analysis if not this indicated the presence of bad data or anomalies.
* Post scaling we can see the difference in the min, max, mean and SD values of the variables of the data.
* **Checking of outliers before scaling:-**



* **Checking for outliers after scaling:-**



* Checking for scaled data post outlier treatment:-



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

Ans:-

Performed PCA and got till seven principal components. Below is the matrix for the same:-

array([[ 4.37712296e-01, 4.29304229e-01, 4.43758174e-01,

9.75222020e-02, 1.07449930e-01, 4.60619821e-01,

3.23799620e-01, -1.76643015e-02, 3.03534281e-02,

8.82769536e-02, 1.13123873e-01, 1.60382927e-01,

1.73041197e-01, 6.32889611e-02, -6.19120065e-02,

6.07953059e-02, 9.37226817e-03],

[ 6.91655847e-02, 8.02995413e-03, -6.73272398e-02,

3.77131032e-01, 3.10873776e-01, -1.15654549e-01,

-2.68902969e-01, 3.49075100e-01, 2.35857271e-01,

3.58691428e-02, -1.50330670e-01, 2.34601058e-01,

2.70248987e-01, -2.29451174e-01, 2.77467003e-01,

4.02858146e-01, 2.11407105e-01],

[-9.86754390e-02, -1.33207948e-01, -6.46596827e-02,

1.56493566e-01, 1.34978356e-01, -4.17163146e-02,

-4.32898611e-02, -9.93815467e-02, -8.15476346e-02,

7.17824204e-01, 5.85151514e-01, -4.79765784e-03,

1.55907082e-03, -9.22273228e-02, -9.10106949e-02,

9.51887519e-02, -1.48116532e-01],

[-1.44069481e-01, -1.14242537e-01, -1.73518128e-01,

-2.96006364e-01, -2.71762231e-01, -1.04189551e-01,

5.81859310e-01, 1.79146731e-01, 4.16544701e-01,

7.83737840e-02, -6.86584797e-03, 1.78213329e-01,

2.81053771e-01, -1.29649421e-01, -1.49847990e-01,

2.27877277e-01, -1.07044857e-01],

[ 2.42235395e-01, 2.09948708e-01, 6.13383277e-02,

-1.50387831e-01, -1.97743926e-01, -5.47951030e-03,

-1.03462040e-01, 2.27509157e-01, 2.90591377e-01,

1.66208341e-01, -2.29313278e-02, -4.70515314e-01,

-5.30729939e-01, -2.95846433e-01, -3.86103570e-02,

2.38197432e-01, 8.86563235e-02],

[-1.04439437e-01, -1.12114125e-01, 6.63333930e-04,

2.69406091e-01, 2.12344395e-01, 6.83352208e-03,

3.48524128e-01, 2.79016332e-02, -1.17630914e-01,

-5.68784111e-01, 4.56321077e-01, -1.35595789e-01,

-2.14688580e-01, -2.94466112e-01, 2.70529004e-02,

1.81769218e-01, -7.75719743e-02],

[-6.63768704e-02, -1.25669852e-01, -8.47685389e-02,

3.32891754e-01, 3.62928802e-01, -4.02708978e-02,

5.19270402e-01, -5.98719043e-02, 6.88066599e-02,

2.78628167e-01, -4.32040353e-01, -2.05709054e-01,

-2.58341756e-01, 1.56586116e-01, 4.88462411e-02,

-1.63291892e-01, 1.53558305e-01]])

* 1. **Extract the eigenvalues, and eigenvectors.**

Ans:-

Below are the extracted eigenvalues and eigenvectors: -

**Eigen Vectors**

%s [[ 4.37712296e-01 -6.91655847e-02 9.86754390e-02 1.44069481e-01

-2.42235395e-01 1.04439437e-01 -6.63768704e-02 -2.93002210e-02

9.46359240e-02 -5.60354443e-03 -1.99945610e-01 7.30049702e-02

-5.11371433e-01 -5.65201845e-01 5.68089988e-02 -2.15034996e-01

-1.16467700e-01]

[ 4.29304229e-01 -8.02995413e-03 1.33207948e-01 1.14242537e-01

-2.09948708e-01 1.12114125e-01 -1.25669852e-01 -4.65420463e-03

3.06899055e-02 -1.81823706e-03 4.34365932e-01 3.05474775e-03

-3.33955269e-01 6.36051011e-01 -3.81035874e-02 7.17407193e-02

-7.82163929e-03]

[ 4.43758174e-01 6.73272398e-02 6.46596827e-02 1.73518128e-01

-6.13383277e-02 -6.63333930e-04 -8.47685389e-02 3.86767760e-02

-1.12323265e-01 -2.40583654e-02 -7.13219873e-01 1.37983947e-02

3.38226871e-01 2.66554019e-01 -1.13958193e-01 1.56120698e-01

8.32764926e-02]

[ 9.75222020e-02 -3.77131032e-01 -1.56493566e-01 2.96006364e-01

1.50387831e-01 -2.69406091e-01 3.32891754e-01 -2.19734939e-01

5.90271288e-02 -8.85424273e-02 9.54421409e-02 1.15156719e-02

-2.21186343e-02 -4.71332685e-02 -6.74038818e-01 -1.87127950e-02

6.24503321e-02]

[ 1.07449930e-01 -3.10873776e-01 -1.34978356e-01 2.71762231e-01

1.97743926e-01 -2.12344395e-01 3.62928802e-01 -1.57683425e-01

1.21681645e-01 -6.96659966e-02 -4.93107903e-02 1.16586964e-01

-2.26206509e-02 1.16917735e-01 6.96280447e-01 1.35428665e-01

-7.71445271e-02]

[ 4.60619821e-01 1.15654549e-01 4.17163146e-02 1.04189551e-01

5.47951030e-03 -6.83352208e-03 -4.02708978e-02 7.17266573e-03

-1.34687138e-01 -2.74153105e-02 4.97557826e-01 2.10155339e-02

5.75081049e-01 -3.75349944e-01 9.50782769e-02 7.74829676e-02

8.09989336e-02]

[ 3.23799620e-01 2.68902969e-01 4.32898611e-02 -5.81859310e-01

1.03462040e-01 -3.48524128e-01 5.19270402e-01 1.80567756e-01

-9.03058898e-02 8.16338527e-02 -2.50367934e-02 -8.78594901e-02

-1.52215404e-01 2.16405566e-02 -2.99277109e-02 -1.39553451e-02

-9.85751153e-03]

[-1.76643015e-02 -3.49075100e-01 9.93815467e-02 -1.79146731e-01

-2.27509157e-01 -2.79016332e-02 -5.98719043e-02 8.38834425e-02

1.25844273e-01 -1.11008501e-01 -9.47543595e-03 -5.15326527e-01

-7.51284850e-02 -1.62325542e-01 3.29232237e-02 6.05924008e-01

2.80484786e-01]

[ 3.03534281e-02 -2.35857271e-01 8.15476346e-02 -4.16544701e-01

-2.90591377e-01 1.17630914e-01 6.88066599e-02 -5.33437471e-02

5.10503940e-01 -3.04256720e-01 -1.46507286e-02 4.77519016e-01

2.49702063e-01 5.20795355e-02 -4.09922756e-02 -8.79015508e-02

3.91700648e-02]

[ 8.82769536e-02 -3.58691428e-02 -7.17824204e-01 -7.83737840e-02

-1.66208341e-01 5.68784111e-01 2.78628167e-01 8.58300850e-02

-1.50751072e-01 6.34796752e-02 -5.63996842e-03 -3.64775244e-02

-2.00312036e-03 1.03279524e-02 -9.54160028e-03 4.55320946e-02

5.45219445e-02]

[ 1.13123873e-01 1.50330670e-01 -5.85151514e-01 6.86584797e-03

2.29313278e-02 -4.56321077e-01 -4.32040353e-01 2.50049293e-01

4.01469382e-01 -1.38842567e-02 1.62635768e-02 -3.13370219e-02

-3.09062095e-02 5.11344908e-03 1.05150007e-03 -5.42587945e-03

-2.42305989e-02]

[ 1.60382927e-01 -2.34601058e-01 4.79765784e-03 -1.78213329e-01

4.70515314e-01 1.35595789e-01 -2.05709054e-01 -3.30343764e-02

1.32443629e-02 1.21818578e-01 -2.27302386e-02 -9.88486889e-03

-8.23518100e-02 3.30925315e-02 8.95034424e-02 -3.27935558e-01

6.82069793e-01]

[ 1.73041197e-01 -2.70248987e-01 -1.55907082e-03 -2.81053771e-01

5.30729939e-01 2.14688580e-01 -2.58341756e-01 -3.05033469e-04

-2.81289382e-02 7.60045454e-02 -5.85227975e-03 1.07174890e-01

-2.84362688e-02 -4.06119291e-02 -1.08433373e-01 3.05222074e-01

-5.45946582e-01]

[ 6.32889611e-02 2.29451174e-01 9.22273228e-02 1.29649421e-01

2.95846433e-01 2.94466112e-01 1.56586116e-01 8.19862726e-02

4.01554974e-01 -4.74422131e-01 -1.34246309e-02 -4.99538886e-01

6.45426129e-02 4.22966468e-02 -1.00467843e-02 -2.22456957e-01

-1.44812240e-01]

[-6.19120065e-02 -2.77467003e-01 9.10106949e-02 1.49847990e-01

3.86103570e-02 -2.70529004e-02 4.88462411e-02 8.33979688e-01

-2.48846170e-01 -2.88865506e-01 3.70166573e-02 1.82765261e-01

-3.46527356e-02 1.30644574e-03 -8.67006910e-03 -8.25275627e-02

9.01929323e-03]

[ 6.07953059e-02 -4.02858146e-01 -9.51887519e-02 -2.27877277e-01

-2.38197432e-01 -1.81769218e-01 -1.63291892e-01 -1.55162228e-01

-3.18386901e-01 -1.29420572e-01 -1.41993134e-02 -3.85010283e-01

1.77386458e-01 1.11856325e-01 1.00512818e-01 -4.85742428e-01

-2.75539293e-01]

[ 9.37226817e-03 -2.11407105e-01 1.48116532e-01 1.07044857e-01

-8.86563235e-02 7.75719743e-02 1.53558305e-01 2.90923163e-01

3.83453994e-01 7.24067974e-01 5.50333325e-03 -1.84191991e-01

2.10498159e-01 3.80273057e-02 -1.89736224e-02 -1.71554799e-01

-1.37525715e-01]]

**Eigen Values**

%s [0.44570119 0.27042032 0.05556861 0.05147092 0.04706552 0.03580954

0.0337879 0.02558063 0.02338909 0.01396077 0.00213739 0.01106955

0.00971056 0.00349932 0.00399469 0.00717635 0.00653673]

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

Ans:-

**Eigenvector of 1st PC:-**

4.37712296e-01\* Apps,

4.29304229e-01\* Accept

4.43758174e-01\* Enroll

9.75222020e-02\* Top10perc

1.07449930e-01\* Top25perc

4.60619821e-01\* F.Undergrad

3.23799620e-01\* P.Undergrad

-1.76643015e-02\* Outstate

3.03534281e-02\* Room.Board

8.82769536e-02\* Books

1.13123873e-01\* Personal

1.60382927e-01\* PhD

1.73041197e-01\* Terminal

6.32889611e-02\* S.F.Ratio

-6.19120065e-02\* perc.alumni

6.07953059e-02\* Expend

9.37226817e-03\* 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?**

Ans:-

Below is the cumulative values of Eigenvalues:-

Cumulative Variance Explained [ 42.57427604 68.40537055 73.71339671 78.63000314 83.12579642

86.54639554 89.773884 92.21739719 94.45156979 95.7851308

96.84251659 97.77008847 98.45558772 99.07998907 99.46157036

99.79583225 100.

The cumulative value of is the cumulative sum of compression across variables. The cumulative value of eigenvalue shows the amount of compressed information present at each variable.

eigenvalues give you the factors by which this compression occurs.

In this case we can see that at 7th variable se get 89.77% (approx. 90%) of data and post 7th variable the impact is not much on the data.

Eigenvectors represent directions.

**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]**

Ans:-

• We see 17 variables need to be addressed, which requires use of PCA (Dimension Reductionality)

• After PCA for this case study we found that the 90% confidence achieved by first seven PCA components only, which is very useful compare to 17 different variables.

•PCA becomes useful by dimension reduction feature by covering 90% of confidence in only 7 features.