2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

* Based on the information on the dataset except for Names every other dimension are numeric ones.
* Based on the percentage of unique value counts for each of the integer variable in the dataset in combination with the data dictionary definitions we can safely assume columns other than Names as continues variables
* There are no missing values across any of the variables
* Also considering all independent variables are integers we can rule out presence of junk characters or unusual values that needs to be imputed for PCA.

2.1 Part A: **Univariate analysis:**

Outliers:

Except for Top25percent every other variable has high amount of outliers with outstate having minimal outlier with positive skewness.

Distribution:

Top25percent while not perfectly normally distributed is somewhat closer to it while rest of the variables are mostly

Skewness:

All of the variables indicates skewness. Terminal, PhD and Grad.Rate being left skewed while rest of the variables are positively skewed.

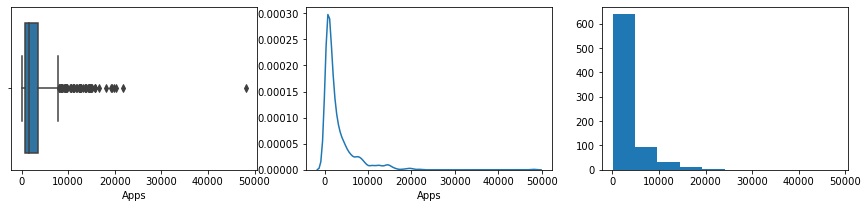
Following depicts the univariate analysis for each of the continuous variables in the dataset across boxplot, kernel density estimate diagram and histograms along with comments on if they have outliers and if they are normally distributed.

Shapiro test has been performed in each of the variable to test the normal distribution.

1. Univariate analysis for Apps

Mean is 3001.638353, Median is 1558.000000, Modes are [3]

Column Apps has outliers

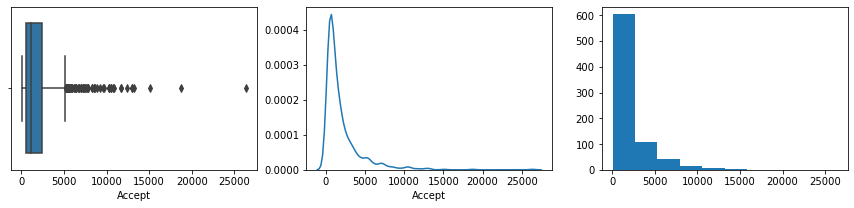


Column Apps is not normally distributed

2. Univariate analysis for Accept

Mean is 2018.804376, Median is 1110.000000, Modes are [4]

Column Accept has outliers

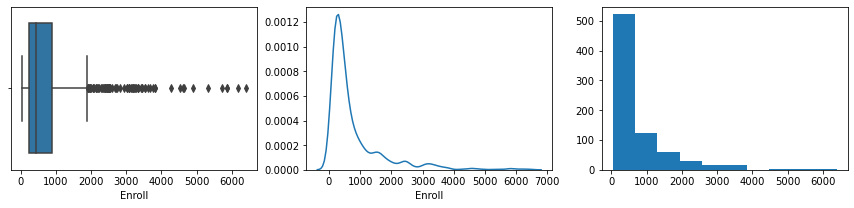


Column Accept is not normally distributed

3. Univariate analysis for Enroll

Mean is 779.972973, Median is 434.000000, Modes are [5]

Column Enroll has outliers

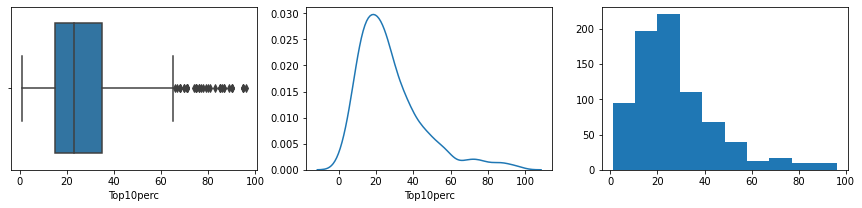


Column Enroll is not normally distributed

4. Univariate analysis for Top10perc

Mean is 27.558559, Median is 23.000000, Modes are [37]

Column Top10perc has outliers

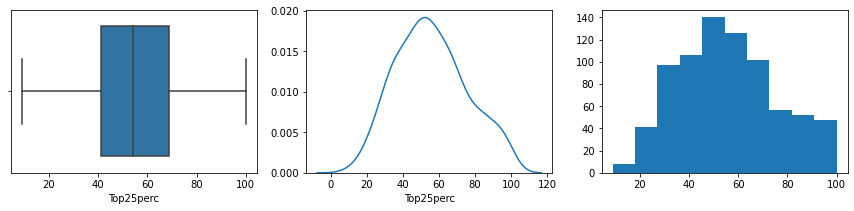


Column Top10perc is not normally distributed

5. Univariate analysis for Top25perc

Mean is 55.796654, Median is 54.000000, Modes are [20]

Column Top25perc does not have outliers

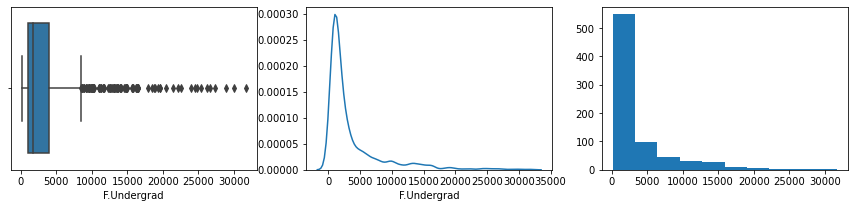


Column Top25perc is not normally distributed

6. Univariate analysis for F.Undergrad

Mean is 3699.907336, Median is 1707.000000, Modes are [3]

Column F.Undergrad has outliers

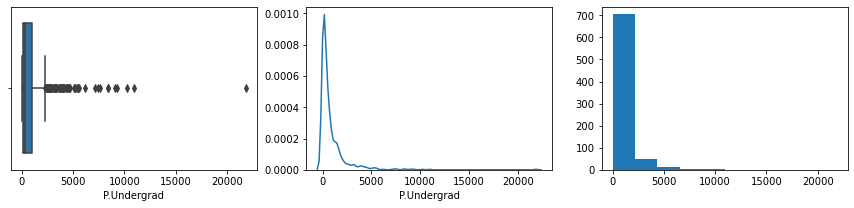


Column F.Undergrad is not normally distributed

7. Univariate analysis for P.Undergrad

Mean is 855.298584, Median is 353.000000, Modes are [7]

Column P.Undergrad has outliers

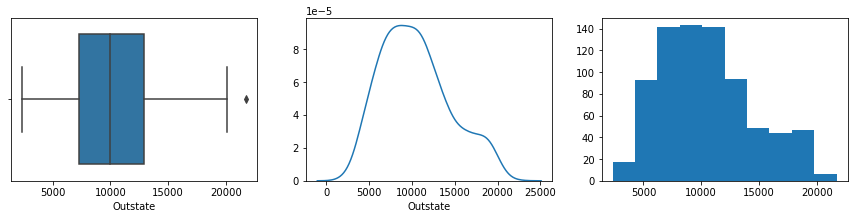


Column P.Undergrad is not normally distributed

8. Univariate analysis for Outstate

Mean is 10440.669241, Median is 9990.000000, Modes are [13]

Column Outstate has outliers

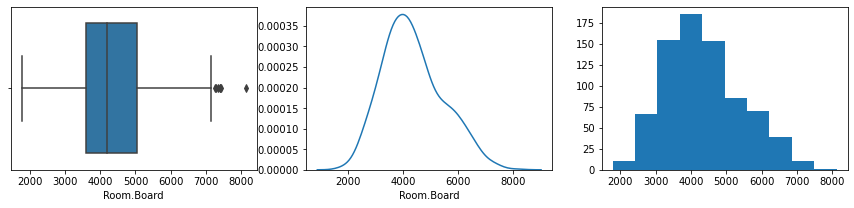


Column Outstate is not normally distributed

9. Univariate analysis for Room.Board

Mean is 4357.526384, Median is 4200.000000, Modes are [9]

Column Room.Board has outliers

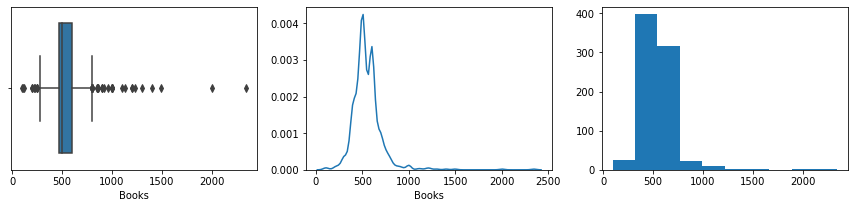


Column Room.Board is not normally distributed

10. Univariate analysis for Books

Mean is 549.380952, Median is 500.000000, Modes are [178]

Column Books has outliers

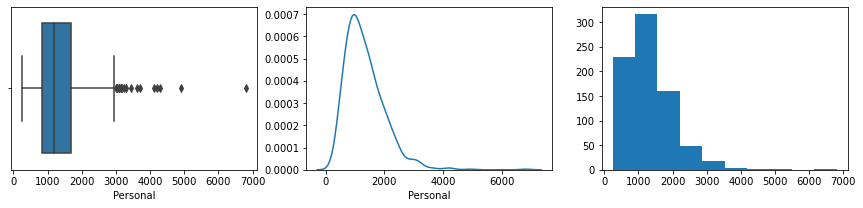


Column Books is not normally distributed

11. Univariate analysis for Personal

Mean is 1340.642214, Median is 1200.000000, Modes are [45]

Column Personal has outliers

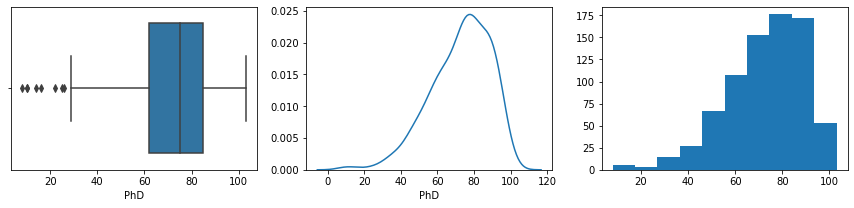


Column Personal is not normally distributed

12. Univariate analysis for PhD

Mean is 72.660232, Median is 75.000000, Modes are [26]

Column PhD has outliers

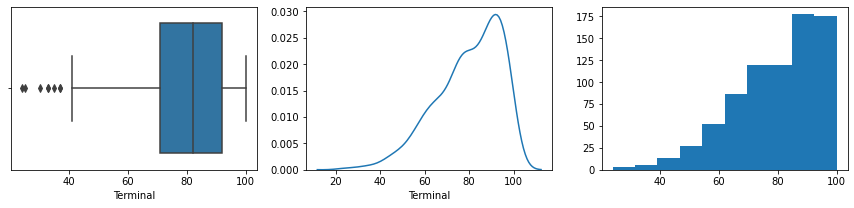


Column PhD is not normally distributed

13. Univariate analysis for Terminal

Mean is 79.702703, Median is 82.000000, Modes are [30]

Column Terminal has outliers

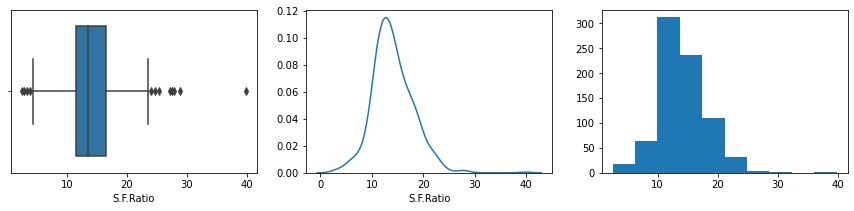


Column Terminal is not normally distributed

14. Univariate analysis for S.F.Ratio

Mean is 14.089704, Median is 13.600000, Modes are [15]

Column S.F.Ratio has outliers

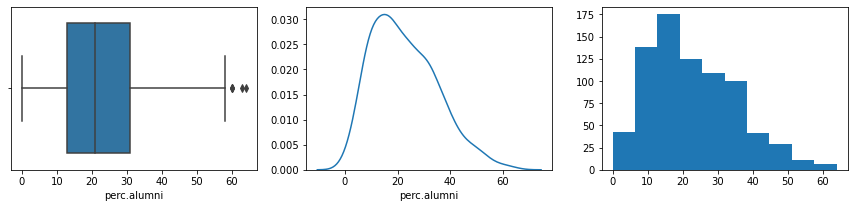


Column S.F.Ratio is not normally distributed

15. Univariate analysis for perc.alumni

Mean is 22.743887, Median is 21.000000, Modes are [32]

Column perc.alumni has outliers

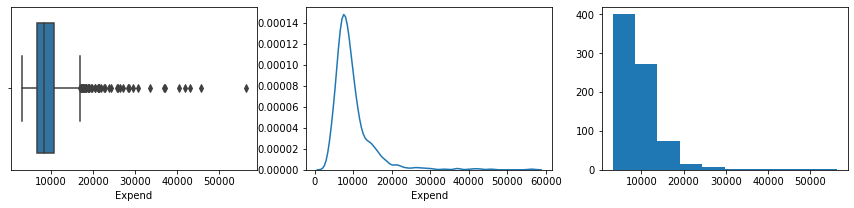


Column perc.alumni is not normally distributed

16. Univariate analysis for Expend

Mean is 9660.171171, Median is 8377.000000, Modes are [2]

Column Expend has outliers

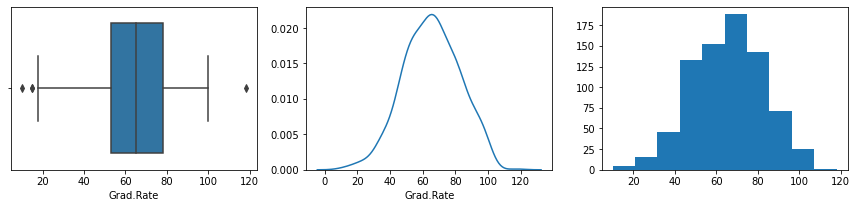


Column Expend is not normally distributed

17. Univariate analysis for Grad.Rate

Mean is 65.463320, Median is 65.000000, Modes are [24]

Column Grad.Rate has outliers



Column Grad.Rate is not normally distributed

2.1 Part B: **Multi variate analysis**

*Multivariate observations based on the pair plot pasted below*

*While the pair plot displays the scatter plots for all combinations of pairs of columns from the original data set aiding to visually observe relative the degree of correlation between the variables, the actual magnitude of co relation can be found coefficient of co relation between the variables.*

​

Based on the coefficient of correlation following observations are being made:

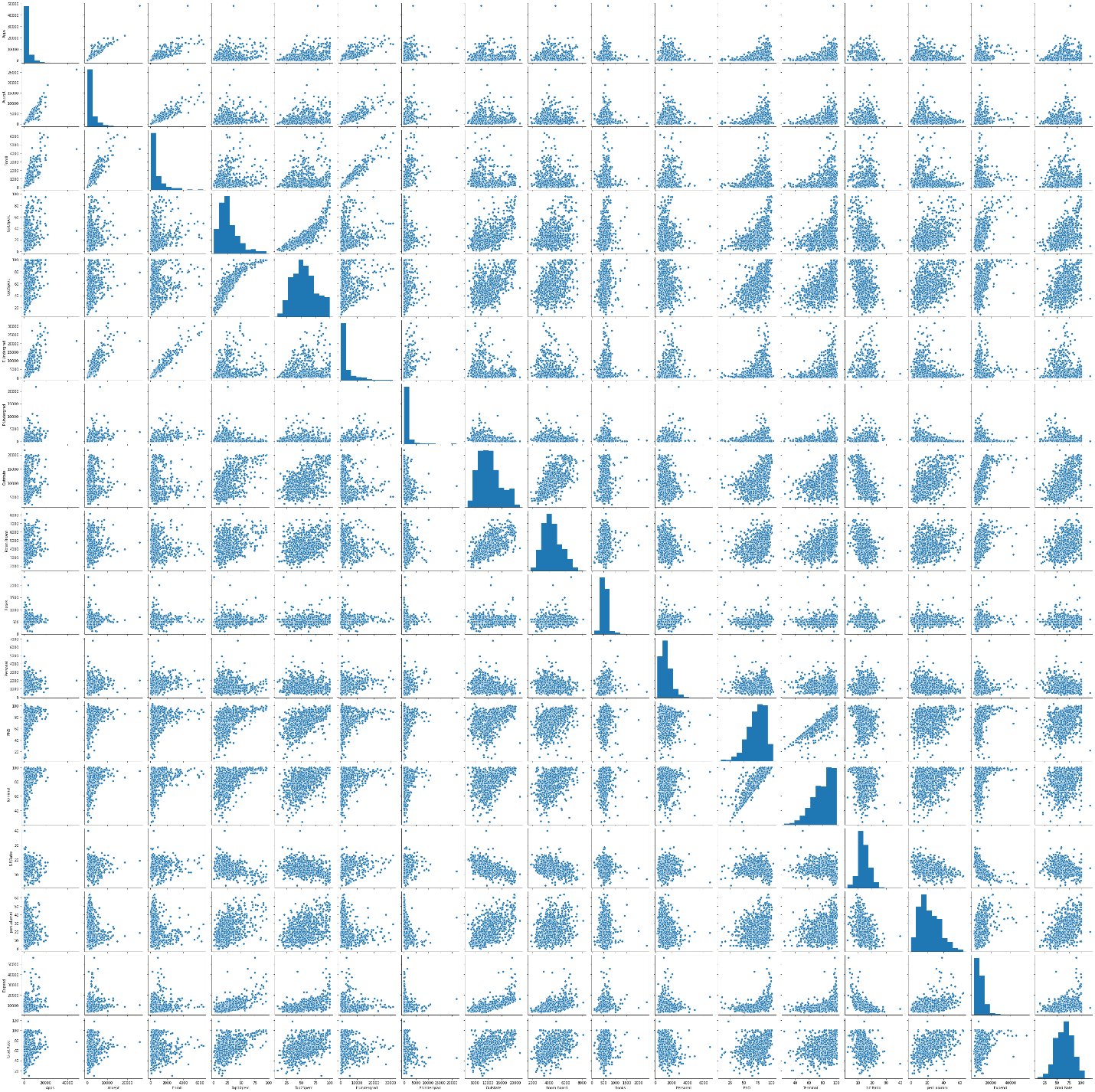
* *Highly and positively correlated:*
  + *Apps with Accept, Enroll and F\_Undergrad*
  + *Accept with Enroll and F.Undergrad*
  + *Top10perc with Top25perc*
  + *Room.Board somewhat with Outstate*
  + *Terminal with PhD*

​

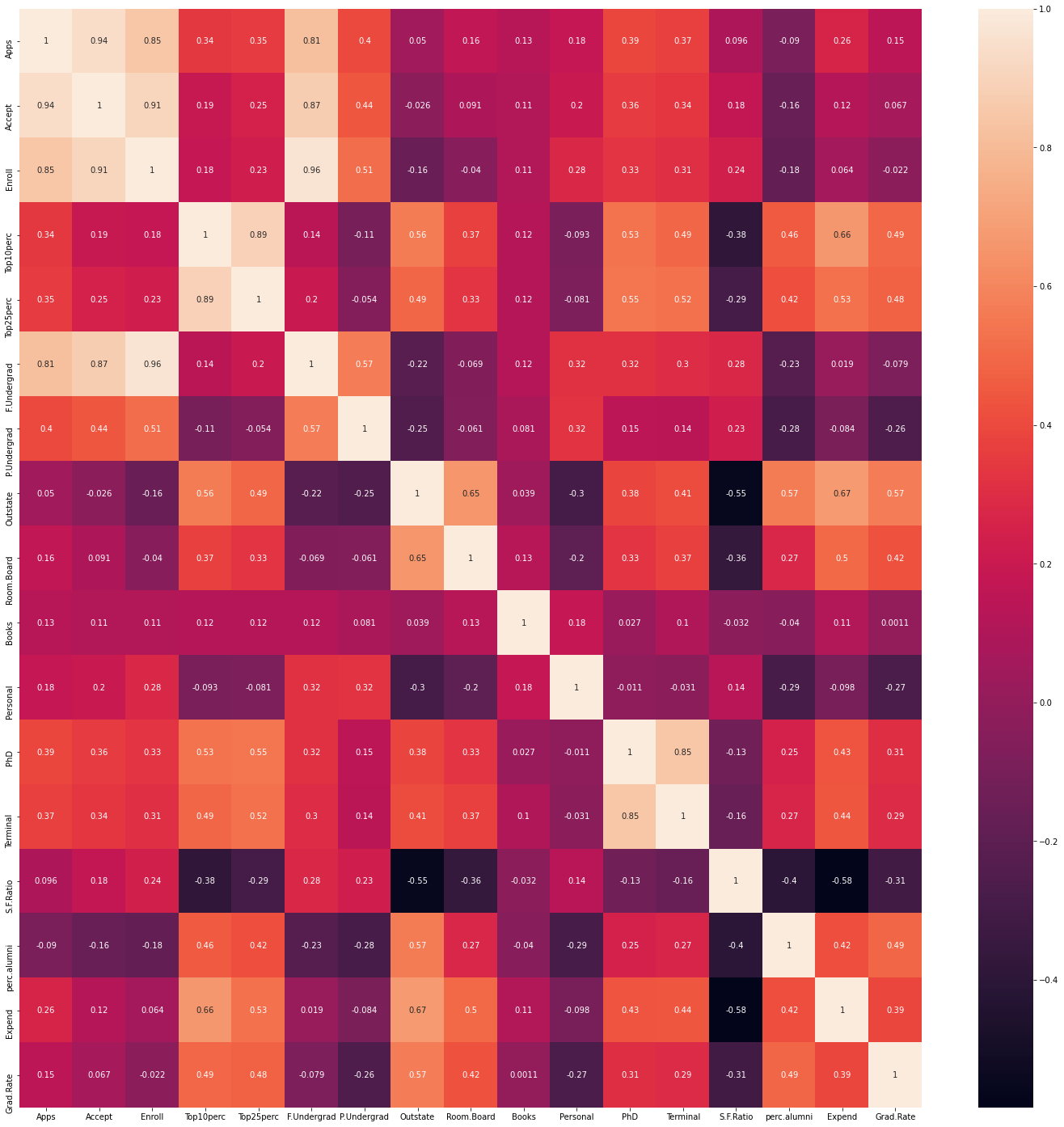
* *Low correlations with positive linearity:*
  + *Books and Personal are comparatively least corelated with every other variable.*

​

* *Negative correlations*
  + *S.F. Ratio has the maximum negative corelation with any variables namely Expend and Outstate*



*The heatmap below reflects the observations narrated above about the correlation with dark shades indicating inverse correlation and light shades indicating positive correlation.*



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

*Inference based on observations for scaling across various scaling options.*

*Scaling options considered: Standardization, Normalization and Logarithmic transformation*

*Observations:*

*1. Given the significant outliers in the original data continues to be retained post to the scaling using Minmaxscaler and standardscaler these methods does not seem to scale the dimension efficiently.*

*2. Logarithmic scaler results in improved distribution of the dataset along with outliers ending up within either side of the whiskers for the good part of it. Almost 66% of the outliers has been reduced as a result of logarithmic scaling.*

Scaling option finalized accordingly will be logarithmic transformation.

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

*Comparison between Covariance and Correlation:*

*1.Covariance is the variance measured among the dimension which denotes the direction of relationship between two independent variables. .Covariance expresses a dimensions variance with itself as well as with other dimensions in the form of matrix. However when it comes to corelation apart from just the direction of relationship it also denotes the measure of strength of relationship between two independent variables.*

*2.Covariance are influenced by unit of variables in the original dataset while corelation which is derived by dividing product of variance between two independent variables with product of their standard deviation has values standardized between -1 and +1 which will be its range.*

*3.Hence coreleation is a unit free standardized measure with a known range of values while covariance can range between infinite* *values on either side of zero.*

*4.Due to robustness of Corelation measure between two variables it is more preferred than covariance for analysis.*

*5.On the other hand, covariance matrix of standardized variable is actually a correlation matrix*

*6.The sign of elements of covariance matrix and the equivalent co relation matrix is identical.corelation is a function of the covariance.*

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

*1. Variables other than Books, Expend, Top25perc, Top10perc, PhD, Terminal and S.F.Ration are significantly treated for their outliers automatically post scaling.*

*2. Out of those variables Enroll and F.Undergrad are completely treated for outliers.*

*3. Few observations on those variables that still retain outliers or introduced new outliers as below.*

* *Books, Expend, continue to have good number of outliers despite scaling however they are equally spread across either side of whiskers compared just right side of the whisker in the original dataset.*
* *Top25perc has introduced new outliers while original dataset did not have any outliers*
* *Top10perc, P.Undergrad, Outstate, Room.Board, perc.alumni had outliers beyond maximum whisker in the original dataset. While scaled dataset does not have any of those outliers it has introduced few outliers to the left of minimum whisker.*
* *Personal variable got most of its outlier right of whisker fixed but couple of outliers introduced to the left of whisker.*
* *PhD, Terminal, S.F.Ratio and Grad.Rate ended up with more outliers to the left of whisker than the original dataset.*

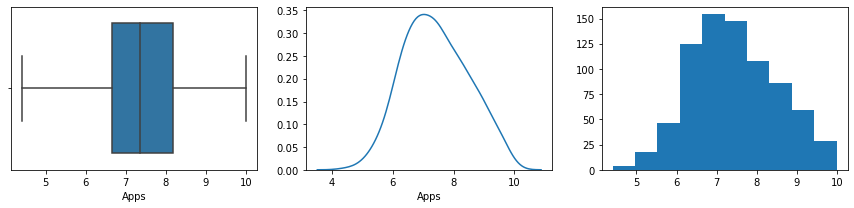
*4. Personal variable is normally distributed post scaling*

*Please find below the analysis done on each of the variables post scaling.*

1. Univariate analysis for Apps

Mean is 7.420076, Median is 7.351158, Modes are [3]

Column Apps does not have outliers

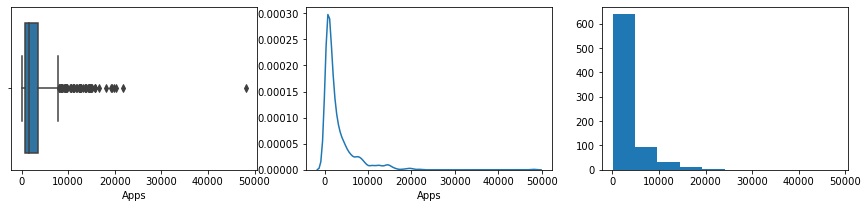
**

Column Apps is not normally distributed

1. Original dataset: Univariate analysis for Apps

Mean is 3001.638353, Median is 1558.000000, Modes are [3]

Column Apps has outliers

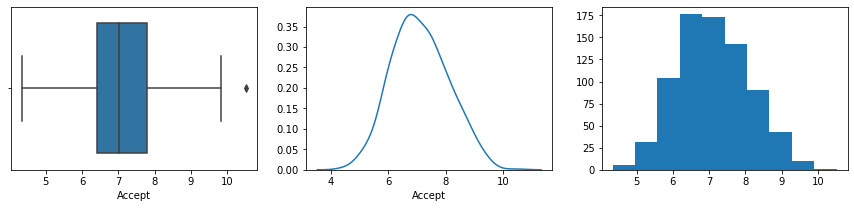
**

Column Apps is not normally distributed

2. Univariate analysis for Accept

Mean is 7.111451, Median is 7.013016, Modes are [4]

Column Accept has outliers

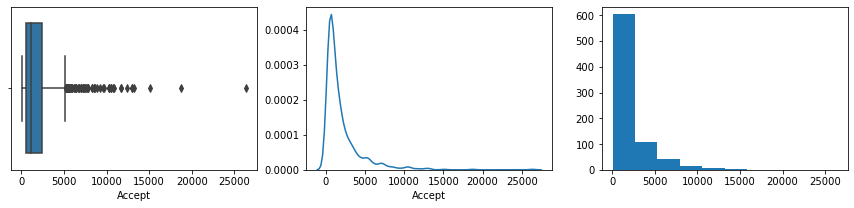
**

Column Accept is not normally distributed

2. Original dataset: Univariate analysis for Accept

Mean is 2018.804376, Median is 1110.000000, Modes are [4]

Column Accept has outliers

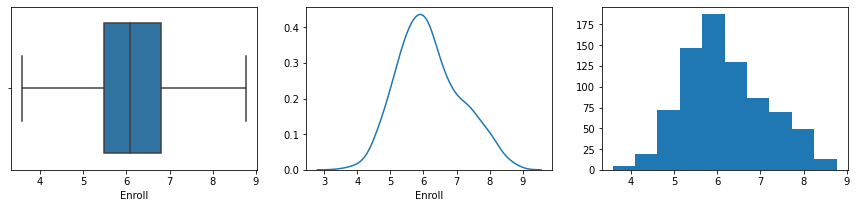
**

Column Accept is not normally distributed

3. Univariate analysis for Enroll

Mean is 6.176126, Median is 6.075346, Modes are [5]

Column Enroll does not have outliers

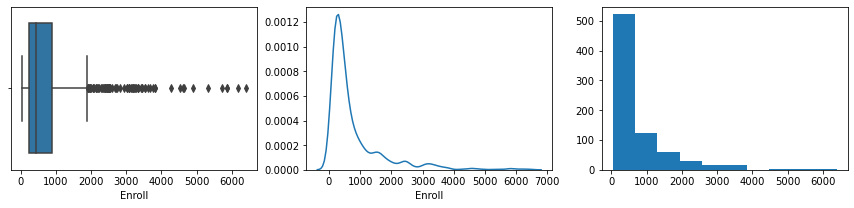
**

Column Enroll is not normally distributed

3. Original dataset: Univariate analysis for Enroll

Mean is 779.972973, Median is 434.000000, Modes are [5]

Column Enroll has outliers

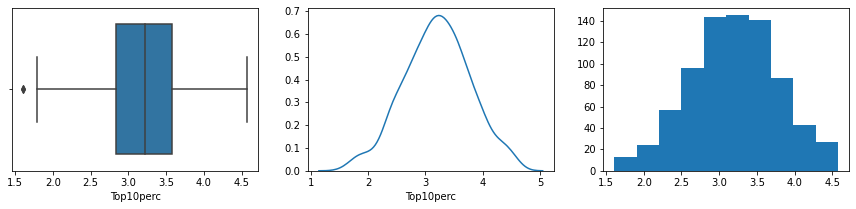
**

Column Enroll is not normally distributed

4. Univariate analysis for Top10perc

Mean is 3.202929, Median is 3.218876, Modes are [37]

Column Top10perc has outliers

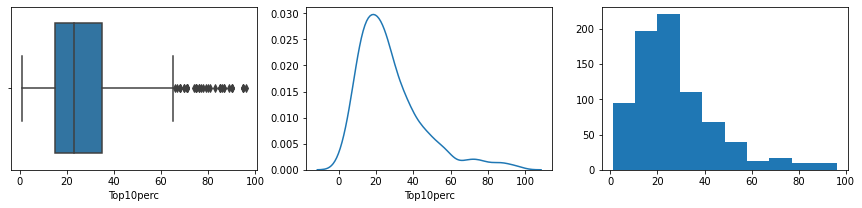
**

Column Top10perc is not normally distributed

4. Original dataset: Univariate analysis for Top10perc

Mean is 27.558559, Median is 23.000000, Modes are [37]

Column Top10perc has outliers

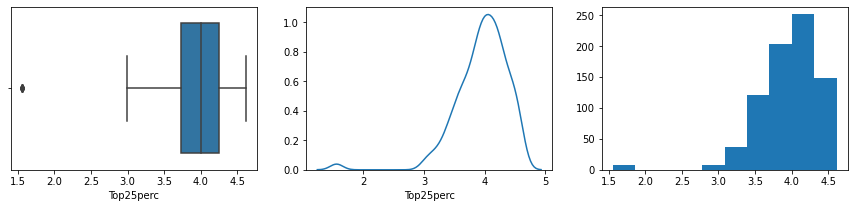
**

Column Top10perc is not normally distributed

5. Univariate analysis for Top25perc

Mean is 3.960345, Median is 4.007333, Modes are [20]

Column Top25perc has outliers

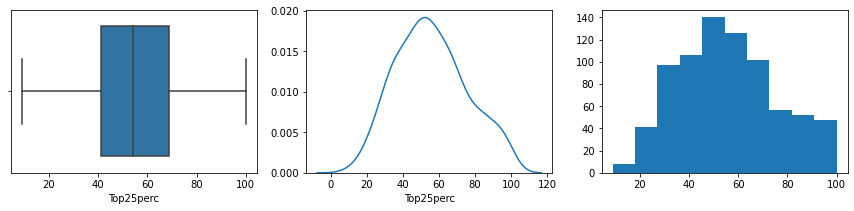
**

Column Top25perc is not normally distributed

5. Original dataset: Univariate analysis for Top25perc

Mean is 55.796654, Median is 54.000000, Modes are [20]

Column Top25perc does not have outliers

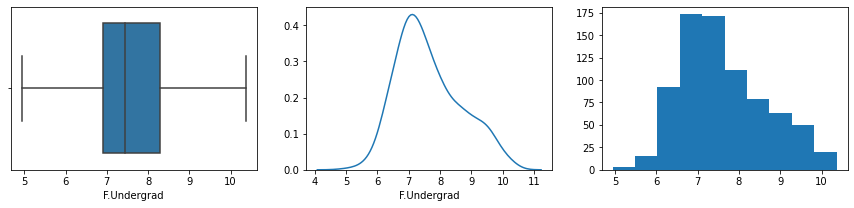
**

Column Top25perc is not normally distributed

6. Univariate analysis for F.Undergrad

Mean is 7.635932, Median is 7.443078, Modes are [3]

Column F.Undergrad does not have outliers

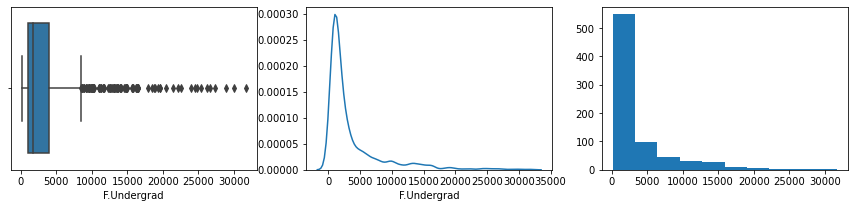
**

Column F.Undergrad is not normally distributed

6. Original dataset: Univariate analysis for F.Undergrad

Mean is 3699.907336, Median is 1707.000000, Modes are [3]

Column F.Undergrad has outliers

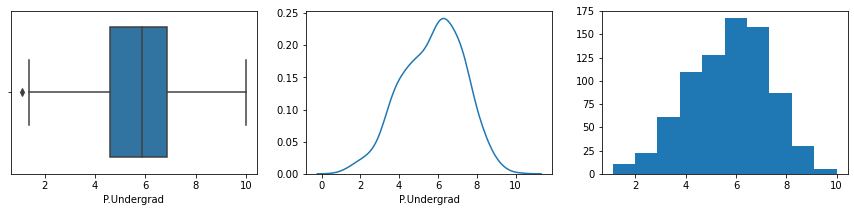
**

Column F.Undergrad is not normally distributed

7. Univariate analysis for P.Undergrad

Mean is 5.728152, Median is 5.869297, Modes are [7]

Column P.Undergrad has outliers

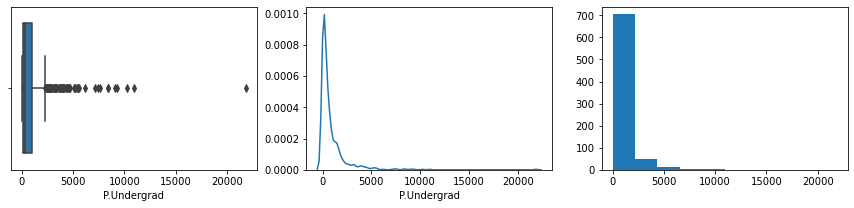
**

Column P.Undergrad is not normally distributed

7. Original dataset: Univariate analysis for P.Undergrad

Mean is 855.298584, Median is 353.000000, Modes are [7]

Column P.Undergrad has outliers

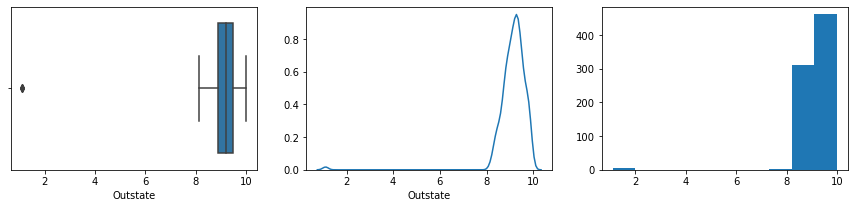
**

Column P.Undergrad is not normally distributed

8. Univariate analysis for Outstate

Mean is 9.140772, Median is 9.209440, Modes are [13]

Column Outstate has outliers

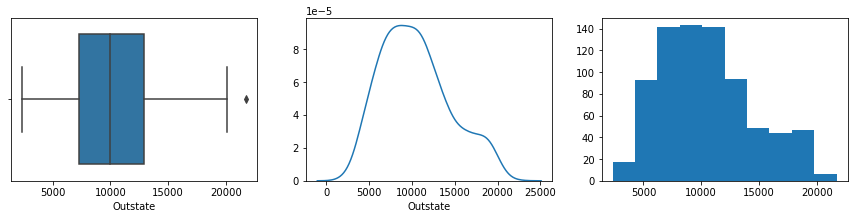
**

Column Outstate is not normally distributed

8. Original dataset: Univariate analysis for Outstate

Mean is 10440.669241, Median is 9990.000000, Modes are [13]

Column Outstate has outliers

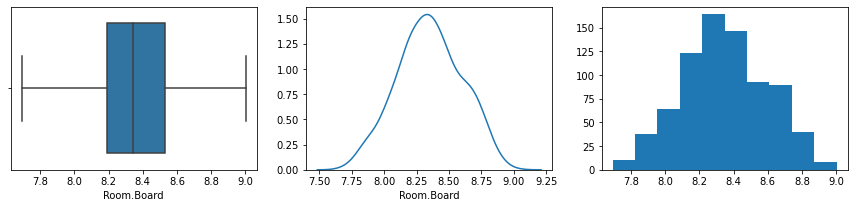
**

Column Outstate is not normally distributed

9. Univariate analysis for Room.Board

Mean is 8.350686, Median is 8.343078, Modes are [9]

Column Room.Board does not have outliers

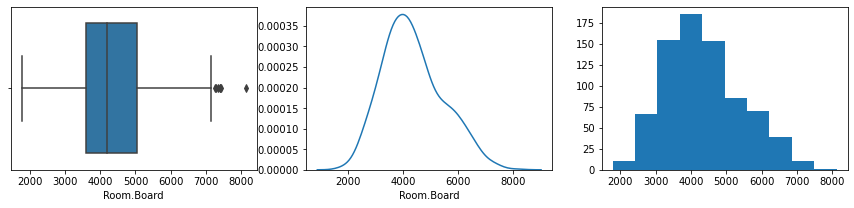
**

Column Room.Board is not normally distributed

9. Original dataset: Univariate analysis for Room.Board

Mean is 4357.526384, Median is 4200.000000, Modes are [9]

Column Room.Board has outliers

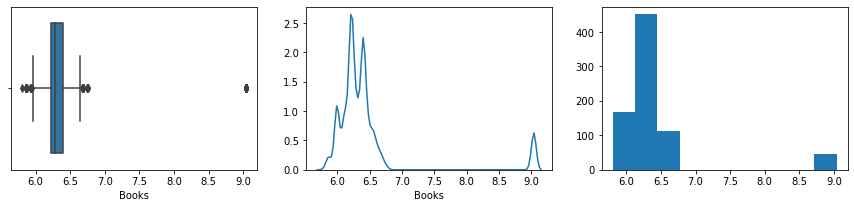
**

Column Room.Board is not normally distributed

10. Univariate analysis for Books

Mean is 6.441340, Median is 6.284134, Modes are [178]

Column Books has outliers

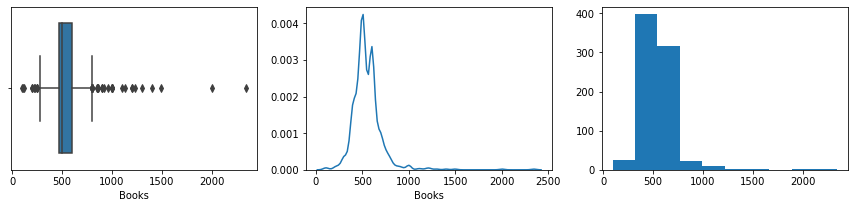
**

Column Books is not normally distributed

10. Original dataset: Univariate analysis for Books

Mean is 549.380952, Median is 500.000000, Modes are [178]

Column Books has outliers

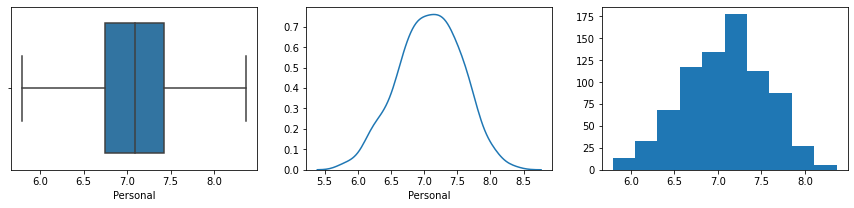
**

Column Books is not normally distributed

11. Univariate analysis for Personal

Mean is 7.081889, Median is 7.090910, Modes are [45]

Column Personal does not have outliers

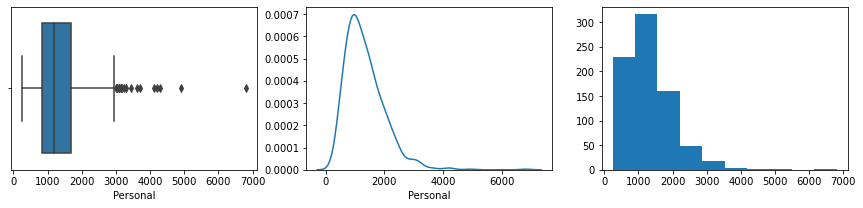
**

Column Personal is not normally distributed

11. Original dataset: Univariate analysis for Personal

Mean is 1340.642214, Median is 1200.000000, Modes are [45]

Column Personal has outliers

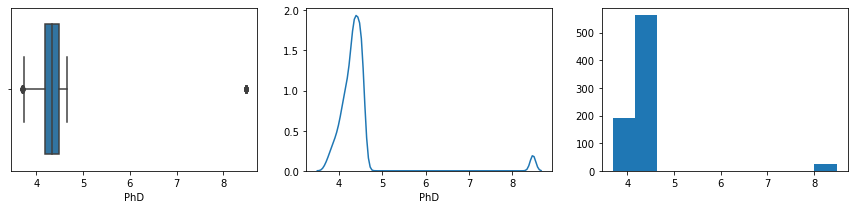
**

Column Personal is not normally distributed

12. Univariate analysis for PhD

Mean is 4.427462, Median is 4.343805, Modes are [26]

Column PhD has outliers

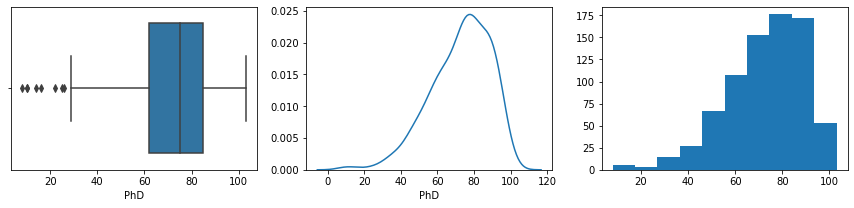
**

Column PhD is not normally distributed

12. Original dataset: Univariate analysis for PhD

Mean is 72.660232, Median is 75.000000, Modes are [26]

Column PhD has outliers

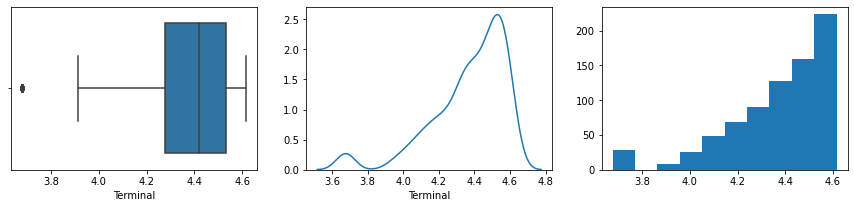
**

Column PhD is not normally distributed

13. Univariate analysis for Terminal

Mean is 4.369178, Median is 4.418841, Modes are [30]

Column Terminal has outliers

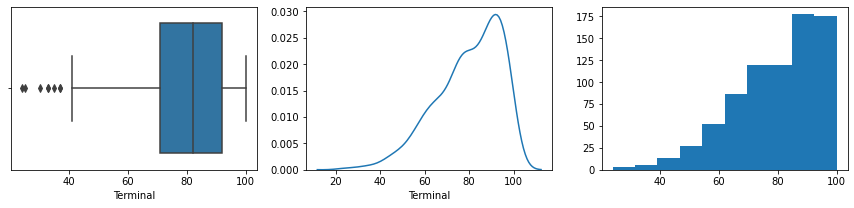
**

Column Terminal is not normally distributed

13. Original dataset: Univariate analysis for Terminal

Mean is 79.702703, Median is 82.000000, Modes are [30]

Column Terminal has outliers

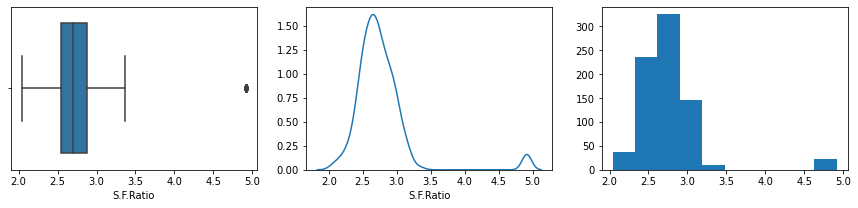
**

Column Terminal is not normally distributed

14. Univariate analysis for S.F.Ratio

Mean is 2.762204, Median is 2.701361, Modes are [22]

Column S.F.Ratio has outliers

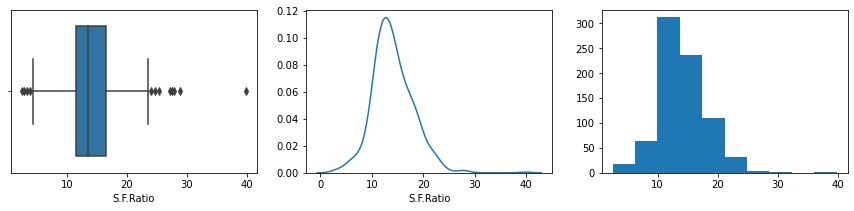
**

Column S.F.Ratio is not normally distributed

14. Original dataset: Univariate analysis for S.F.Ratio

Mean is 14.089704, Median is 13.600000, Modes are [15]

Column S.F.Ratio has outliers

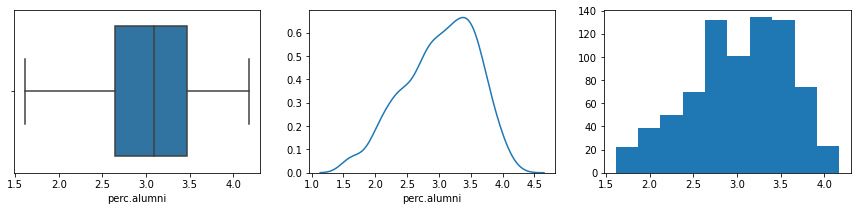
**

Column S.F.Ratio is not normally distributed

15. Univariate analysis for perc.alumni

Mean is 3.023637, Median is 3.091042, Modes are [32]

Column perc.alumni does not have outliers

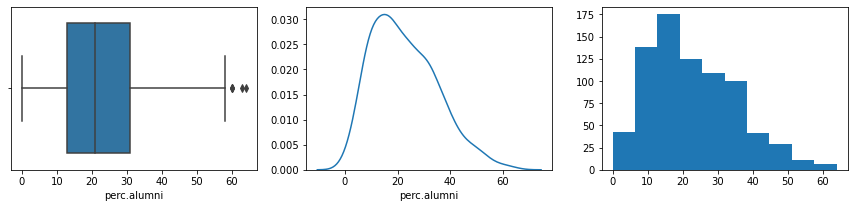
**

Column perc.alumni is not normally distributed

15. Original dataset: Univariate analysis for perc.alumni

Mean is 22.743887, Median is 21.000000, Modes are [32]

Column perc.alumni has outliers

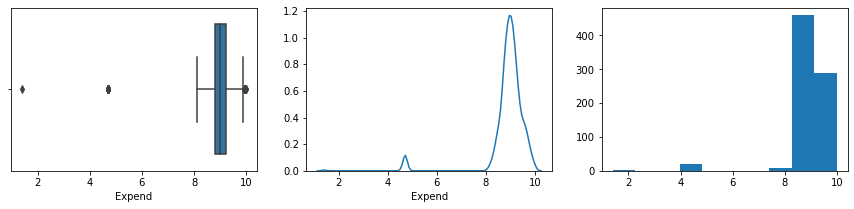
**

Column perc.alumni is not normally distributed

16. Univariate analysis for Expend

Mean is 8.919903, Median is 9.004054, Modes are [21]

Column Expend has outliers

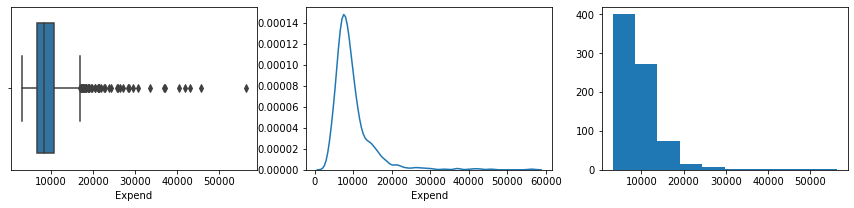
**

Column Expend is not normally distributed

16. Original dataset: Univariate analysis for Expend

Mean is 9660.171171, Median is 8377.000000, Modes are [2]

Column Expend has outliers

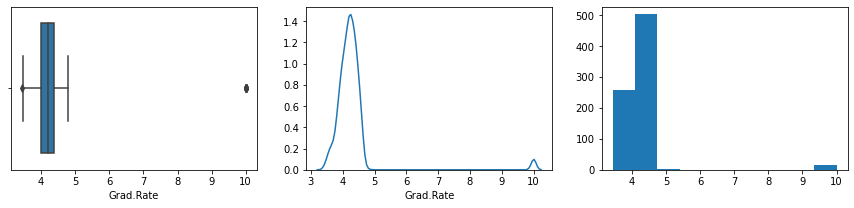
**

Column Expend is not normally distributed

17. Univariate analysis for Grad.Rate

Mean is 4.291241, Median is 4.204693, Modes are [24]

Column Grad.Rate has outliers

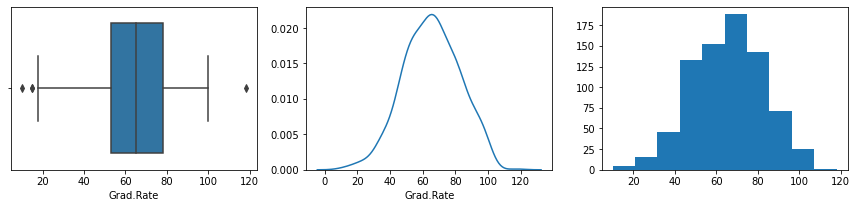
**

Column Grad.Rate is not normally distributed

17. Original dataset: Univariate analysis for Grad.Rate

Mean is 65.463320, Median is 65.000000, Modes are [24]

Column Grad.Rate has outliers

**

Column Grad.Rate is not normally distributed

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

Covariance matrix

[[ 1.00128866 0.95541955 0.91262543 0.32523071 0.32626151 0.87439233

0.37895685 0.01277525 0.18424231 0.05770185 0.19192928 0.01000893

0.45265804 0.25188199 -0.06942519 -0.10662031 -0.00853013]

[ 0.95541955 1.00128866 0.94871582 0.25983855 0.27964596 0.90664593

0.43367266 -0.0099024 0.13195219 0.04534186 0.21298526 -0.00103516

0.43002043 0.24795573 -0.10025888 -0.0555676 -0.03255256]

[ 0.91262543 0.94871582 1.00128866 0.21786442 0.24810373 0.96294356

0.50145483 -0.12729824 -0.0105333 0.05610706 0.28258486 -0.00752911

0.3754636 0.27164187 -0.17069922 -0.10579141 -0.03251622]

[ 0.32523071 0.25983855 0.21786442 1.00128866 0.73648415 0.16428874

-0.26404496 0.30867023 0.33552143 -0.04263824 -0.12274647 -0.00587091

0.43808623 0.01983607 0.42355493 -0.01234464 0.02152219]

[ 0.32626151 0.27964596 0.24810373 0.73648415 1.00128866 0.22379513

-0.13991007 0.2406126 0.29150266 -0.00429549 -0.06341502 -0.04900996

0.4744419 -0.00554412 0.36811822 0.01705732 -0.06462958]

[ 0.87439233 0.90664593 0.96294356 0.16428874 0.22379513 1.00128866

0.58854327 -0.15676453 -0.02445247 0.067161 0.30994846 -0.01062726

0.35644072 0.28324972 -0.24048969 -0.12384906 -0.07395962]

[ 0.37895685 0.43367266 0.50145483 -0.26404496 -0.13991007 0.58854327

1.00128866 -0.2519212 -0.07403701 0.01703668 0.32875726 -0.01453353

0.03117041 0.2033387 -0.44139811 -0.04192537 -0.06854909]

[ 0.01277525 -0.0099024 -0.12729824 0.30867023 0.2406126 -0.15676453

-0.2519212 1.00128866 0.43515264 -0.0363087 -0.17736377 0.05952298

0.17614089 -0.10801639 0.30250703 0.06556907 0.05252247]

[ 0.18424231 0.13195219 -0.0105333 0.33552143 0.29150266 -0.02445247

-0.07403701 0.43515264 1.00128866 -0.02949514 -0.21344729 0.07049765

0.34510098 -0.01673382 0.2714961 0.05867713 0.02022646]

[ 0.05770185 0.04534186 0.05610706 -0.04263824 -0.00429549 0.067161

0.01703668 -0.0363087 -0.02949514 1.00128866 0.04229724 0.14263648

-0.01420618 0.00908497 -0.09644504 0.02998151 -0.04908418]

[ 0.19192928 0.21298526 0.28258486 -0.12274647 -0.06341502 0.30994846

0.32875726 -0.17736377 -0.21344729 0.04229724 1.00128866 -0.00919825

-0.01612421 0.08372049 -0.29104666 -0.05438935 -0.04653986]

[ 0.01000893 -0.00103516 -0.00752911 -0.00587091 -0.04900996 -0.01062726

-0.01453353 0.05952298 0.07049765 0.14263648 -0.00919825 1.00128866

-0.18941157 0.00542516 -0.04878139 0.02743129 0.10286011]

[ 0.45265804 0.43002043 0.3754636 0.43808623 0.4744419 0.35644072

0.03117041 0.17614089 0.34510098 -0.01420618 -0.01612421 -0.18941157

1.00128866 0.04270595 0.25265237 0.06445938 -0.06496179]

[ 0.25188199 0.24795573 0.27164187 0.01983607 -0.00554412 0.28324972

0.2033387 -0.10801639 -0.01673382 0.00908497 0.08372049 0.00542516

0.04270595 1.00128866 -0.13060367 -0.50070988 -0.01847701]

[-0.06942519 -0.10025888 -0.17069922 0.42355493 0.36811822 -0.24048969

-0.44139811 0.30250703 0.2714961 -0.09644504 -0.29104666 -0.04878139

0.25265237 -0.13060367 1.00128866 0.01277898 0.02201253]

[-0.10662031 -0.0555676 -0.10579141 -0.01234464 0.01705732 -0.12384906

-0.04192537 0.06556907 0.05867713 0.02998151 -0.05438935 0.02743129

0.06445938 -0.50070988 0.01277898 1.00128866 -0.02113291]

[-0.00853013 -0.03255256 -0.03251622 0.02152219 -0.06462958 -0.07395962

-0.06854909 0.05252247 0.02022646 -0.04908418 -0.04653986 0.10286011

-0.06496179 -0.01847701 0.02201253 -0.02113291 1.00128866]]

**Unsorted eigen values:**

[4.68339713 3.0926752 1.41509136 1.24276537 1.05460841 0.02274642

0.03946683 0.086803 0.97268673 0.83464394 0.79794695 0.23807285

0.62263808 0.56509328 0.42287993 0.44535779 0.48503395]

**Eigen vectors sorted in the descending order of respective eigen values**

1. Eigen vector for the eigen value 4.683397 as below

[-0.43574848 -0.04033648 -0.02839344 -0.08386651 -0.05582437 -0.1621535

-0.54951212 0.61661101 -0.01288726 0.0823794 0.06674398 -0.0428007

-0.17083491 0.03509236 -0.1655586 0.13363344 0.01228525]

2. Eigen vector for the eigen value 3.092675 as below

[-0.43977136 -0.0032152 -0.07039439 -0.06122482 -0.07042734 0.47624106

0.64779425 0.23835713 0.00333753 0.08525931 0.09304644 -0.07071909

-0.20352865 0.04179608 -0.14628742 0.0461732 0.04503306]

3. Eigen vector for the eigen value 1.415091 as below

[-0.44274238 0.06724214 -0.04818005 -0.01944432 -0.0342385 -0.73602853

0.2704205 -0.35623659 -0.08822306 0.0602057 0.10007782 -0.01409669

-0.1699603 0.01225146 -0.05411991 0.01500162 -0.00181997]

4. Eigen vector for the eigen value 1.242765 as below

[-0.1531815 -0.42767338 0.08352052 0.02601668 0.12061676 0.02927787

0.03300851 0.00825157 -0.27275296 -0.17255393 0.02930615 0.72592816

0.14383001 -0.31198958 -0.12936211 -0.00273623 -0.05146769]

5. Eigen vector for the eigen value 1.054608 as below

[-0.17506509 -0.38970582 0.01245753 0.11281905 0.20442602 -0.0092547

0.02169356 0.03768735 -0.23674019 -0.19441676 0.00455357 -0.63689799

0.27825928 -0.33983298 0.11829606 -0.12517766 -0.19685022]

6. Eigen vector for the eigen value 0.972687 as below

[-0.0602408 -0.33742686 -0.01608734 -0.24308664 -0.16316366 -0.02895097

0.01183726 -0.15274551 0.53707735 0.00057189 -0.14721402 -0.05782372

0.36533447 0.17055996 -0.43925509 0.20330662 -0.24884903]

7. Eigen vector for the eigen value 0.834644 as below

[-0.02788799 0.05354694 -0.13210067 -0.35866925 0.69924122 0.0011677

0.01503737 0.00198278 -0.0422829 0.4113859 -0.41470647 0.03980882

-0.03405823 0.01628391 -0.01208837 -0.07334249 -0.09771531]

8. Eigen vector for the eigen value 0.797947 as below

[-0.13959089 0.23530809 -0.10062291 0.09036058 0.07313049 0.00802271

0.00564092 0.01697376 -0.17147227 -0.67920894 -0.54004086 -0.00453515

0.0078368 0.28857241 -0.17810667 0.01496684 0.05872352]

9. Eigen vector for the eigen value 0.622638 as below

[-0.23679506 -0.28977917 -0.08534826 0.20727459 0.00058569 -0.0085205

0.01553682 0.0098016 0.09287959 0.15641987 -0.18715872 0.07387939

0.29894008 0.33347573 0.5499299 0.22736655 0.42547459]

10. Eigen vector for the eigen value 0.565093 as below

[-0.1578606 0.1124749 0.62631547 -0.05906361 0.08224955 -0.00502452

-0.01779497 -0.00478765 0.15491229 0.03083195 -0.02813789 -0.05152411

0.20909903 -0.0541467 -0.21049461 -0.43910531 0.49736245]

11. Eigen vector for the eigen value 0.485034 as below

[ 0.03187801 -0.01951373 0.0839663 -0.3809468 -0.58477271 0.02090632

-0.0025871 -0.02493959 -0.51675438 0.26807435 -0.37099281 -0.0503148

0.13645654 -0.01805912 0.03254145 -0.06203944 -0.02206309]

12. Eigen vector for the eigen value 0.445358 as below

[ 0.06800833 -0.07758216 -0.72124328 0.01164169 -0.07678735 -0.01266895

-0.05832671 -0.01988239 -0.01613532 0.05396219 0.09671182 -0.0144102

0.14311852 -0.07852939 -0.28138176 -0.41375177 0.41161301]

13. Eigen vector for the eigen value 0.422880 as below

[ 0.06233553 -0.40667055 0.11025774 0.09529335 0.03178267 0.01177129

-0.0366394 -0.02156879 -0.15706427 0.04229109 0.12110333 0.00474049

-0.24916508 0.65179097 -0.03698605 -0.47459905 -0.22774163]

14. Eigen vector for the eigen value 0.238073 as below

[ 0.01569946 0.02288973 -0.03765677 -0.71809048 0.12368662 0.0009883

0.00706678 0.0069796 -0.08660513 -0.36401123 0.44008901 -0.00578583

0.14779024 0.19527671 0.23274715 0.05809664 0.12458636]

15. Eigen vector for the eigen value 0.086803 as below

[ 0.02617649 -0.3391353 -0.01454978 -0.2511262 -0.14735101 -0.02215151

-0.02034081 -0.03849309 0.36465401 -0.21647268 -0.30161455 -0.01497847

-0.56244469 -0.29085778 0.2714619 -0.17701433 0.14117693]

16. Eigen vector for the eigen value 0.039467 as below

[-0.24223912 0.30285015 -0.11740326 -0.00171752 -0.1438405 -0.0369528

0.04352657 0.18654761 0.27566139 -0.04060601 -0.04142778 0.21024408

0.2931258 -0.0513458 0.36780528 -0.48737944 -0.43832625]

17. Eigen vector for the eigen value 0.022746 as below

[-0.43882631 0.10636605 -0.05083581 -0.00507174 -0.01313547 0.4479116

-0.44231902 -0.6107607 -0.03009068 0.03446746 0.08278873 -0.00857632

-0.09652062 -0.02424118 0.03525194 -0.01143406 -0.06113393]

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

Explicit form of first PC in the form of eigen vector is below :

[-0.43574848 -0.04033648 -0.02839344 -0.08386651 -0.05582437 -0.1621535

-0.54951212 0.61661101 -0.01288726 0.0823794 0.06674398 -0.0428007

-0.17083491 0.03509236 -0.1655586 0.13363344 0.01228525]

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.

2.7 Part A

Cumulative values of eigen values and how it helps to decide on the optimum number of principle components:

* Eigen values are the variances mapped to the corresponding eigen vector.
* Highest eigen values based on its proportion to the total eigen values derived out of decomposition of the covariance matrix of the original dataset helps us in optimizing/determining the optimal number of principal components towards prediction analysis. This method is also referred to as dimension reduction.
* Hence cumulative values of eigen values in its descending order helps us to determine optimal number of corresponding eigen vectors to ensure appropriate percentage of coverage of total variance among the data in the original dataset.
* Eigen values assists in how many PCA to pick based on its cumulative percentage coverage compared to the total variance of the original data set.
* Each eigen value derived out of decomposition of the covariance matrix of the original dataset has corresponding eigen vector which is nothing but the principle component for the given eigen value.

2.7 Part B

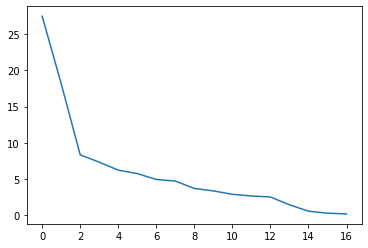
What do eigen vectors indicate?

* Eigen vector establishes a direction that captures most of the variances on its axis and there will be as many eigen vectors until all of the data could be explained in as many dimensions.
* Each eigen vector will correspond to an appropriate eigen value that explains the variance captured for the dataset equivalent to that of a single eigen vector (i.e Principal component analysis).
* Eigen values assists in how many PCA to pick based on its cumulative percentage coverage compared to the total variance of the original data set.
* Eigen vector can be used for loadings of factor analysis.Influence of principal component on a dimension is called weights or "loadings" which is nothing but eigen vector.
* Eigen vector reduces off diagonal elements of covariance metrics from which it is formed to 0. with every eigen vector corresponding to a Principal component this property of the eigen vector ensures orthogonality of every other principal component also enables to overcome the challenge of multi colinearity among variables when it comes to prediction.

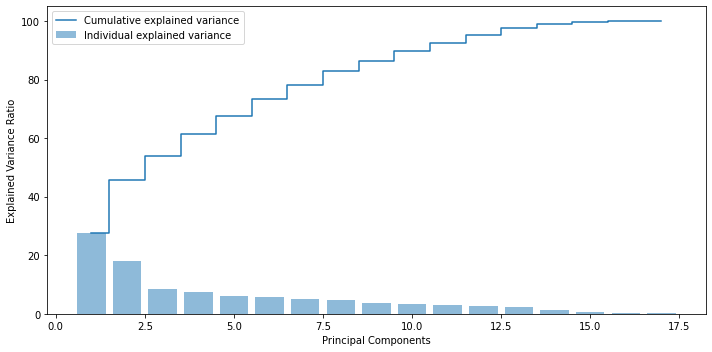
2.7 Part C

Based on the scree plot for variance explained and the step plot for cumulative eigen values it could be narrowed down that for the given data set we could use the top 9 Principle components towards reducing the number of dimensions from 17 to 9 covering almost 86% of the variation in the dataset.

Scree plot as below



Bar plot including the step plot suggesting the cumulative variance explained for the principal component starting from 1 to 17.



Accordingly below data frame has been constructed for the top 9 principle components.

