# BUSINESS REPORT

## ADVANCED STATISTICS

SRINIDHI DEVAN

# BUSINESS REPORT

***Problem 1:***

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments.

***About the Data:*** This dataset contains 4 variables which are as follows:

1) A: One active ingredient of a new compound
2) B: The other active ingredient of a new compound
3) Volunteer: The volunteers in the research (1,2,3,4)
4) Relief: Hours of relief after the nine treatments

Basis this data, we are required to perform ANOVA (Analysis of Variance).

| A | B | Volunteer | Relief |
|---|---|---|---|
| 1 | 1 | 1 | 2.4 |
| 1 | 1 | 2 | 2.7 |
| 1 | 1 | 3 | 2.3 |
| 1 | 1 | 4 | 2.5 |
| 1 | 2 | 1 | 4.6 |

*Assumptions of ANOVA:*

1. *The samples drawn from different populations are independent & random*
2. *The response variables of the population are continuous & normally distributed*
3. *The variances of the population are equal (approximately)*

*are assumed to be satisfied.*

***1.1 State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually. [both statement and statistical form like Ho=mu, Ha>mu]***

**Hypothesis for A:**

*Null:* The mean hours of relief for the different levels of component A are equal

*Alt:* The mean hours of relief for at least one of the levels of component A is (are) not equal

Alternatively,

$$Null: \mu 1 = \mu 2 = \mu 3$$

*Alt:* At least one μ j's is (are) not equal

**Hypothesis for B:**

*Null:* The mean hours of relief for the different levels of component B are equal

*Alt:* The mean hours of relief for at least one of the levels of component B is (are) not equal

Alternatively,

$$Null: \mu 1 = \mu 2 = \mu 3$$

*Alt:* At least one μ j's is (are) not equal

***1.2 Perform one-way ANOVA for variable 'A' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.***

| Column1 | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(A) | 2 | 220.02 | 110.01 | 23.465 | 4.58E-07 |
| Residual | 33 | 154.17 | 4.688 | | |

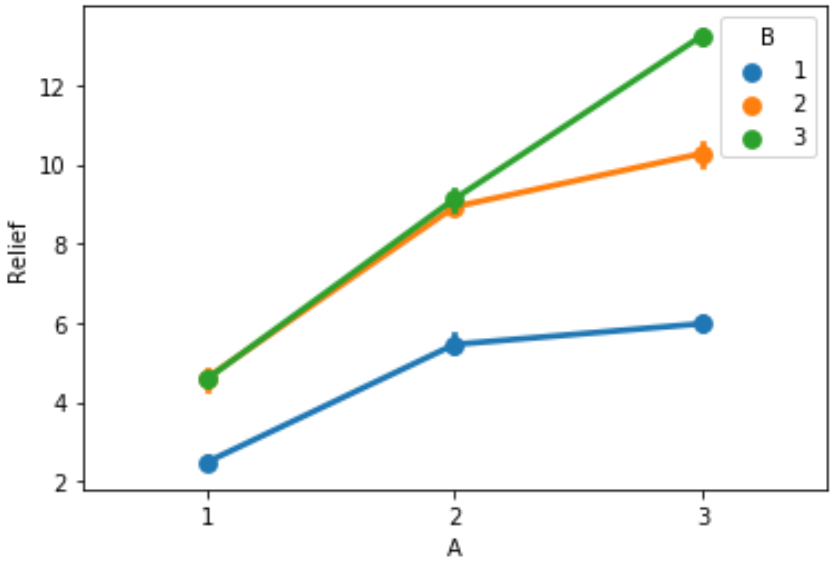
Here, p-Value=0.00000045 < α=0.05, so we reject null and conclude that mean hours of relief for at least one of the levels of component A is (are) not equal.

***1.3 Perform one-way ANOVA for variable 'B' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.***

| Column1 | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(B) | 2 | 123.66 | 61.83 | 8.126777 | 0.00135 |
| Residual | 33 | 251.07 | 7.608182 | | |

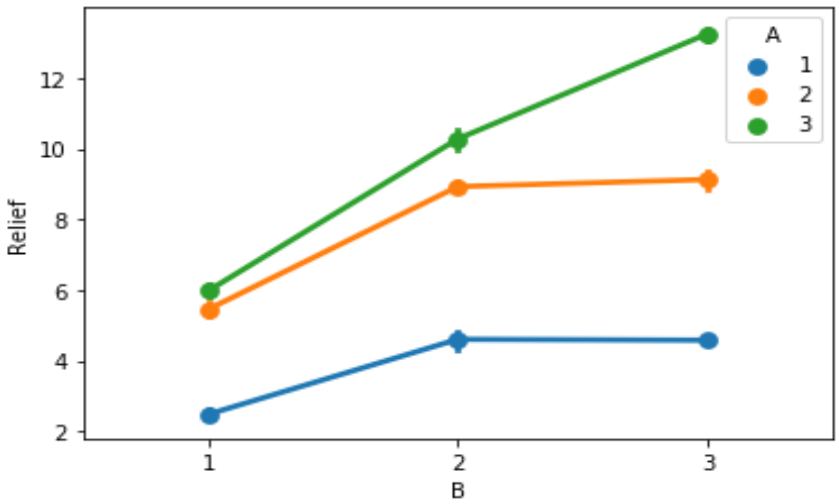
Here, p-Value=0.0005 < α=0.05, so we reject null and conclude that mean hours of relief for at least one of the levels of component B is (are) not equal.

***1.4 Analyse the effects of one variable on another with the help of an interaction plot. What is the interaction between the two treatments?***




Relief (in hours) V. Ingredient A (with hue: Ingredient B)

.




Relief (in hours) V. Ingredient B (with hue: Ingredient A)

From the interaction plot, we can observe that the lines are not parallel to each other implying there is a relationship between the independent variables ('A' & 'B') and dependent variable ('Relief').

Further observing, level 2 & 3 (from both the graphs above) are certainly not parallel implying a (stronger) relationship of levels 2 & 3 with 'Relief.

***1.5 Perform a two-way ANOVA based on the different ingredients (variable 'A' & 'B' along with their interaction 'A*B') with the variable 'Relief' and state your results.***

***Null:*** No interaction between compound A & B

***Alt:*** There is interaction between compound A & B

| Column1 | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(A) | 2 | 220.02 | 110.01 | 1827.858 | 1.51E-29 |
| C(B) | 2 | 123.66 | 61.83 | 1027.329 | 3.35E-26 |
| C(A):C(B) | 4 | 29.425 | 7.35625 | 122.2269 | 6.97E-17 |
| Residual | 27 | 1.625 | 0.060185 | | |

- The p-value for 'A' (= 1.51e-29) which is less than α (= 0.05), and conclude that mean hours of relief for at least one of the levels of component A is (are) not equal
- The p-value for 'B' (= 3.35e-26) which is less than α=0.05 and conclude that mean hours of relief for at least one of the levels of component B is (are) not equal
- Here, p-Value for 'A*B' (= 6.97e-17) is less than α=0.05, so we reject null and conclude that there is interaction (or relationship) between compound A & B.

*Note:*

*The null hypothesis for a main effect is that the response ('Relief') mean for all factor levels are equal.*

*The null hypothesis for an interaction effect is that the response ('Relief') mean for the level of one factor does not depend on the value of the other factor level.*

*If an interaction term is statistically significant, the relationship between a factor and the response differs by the level of the other factor. In this case, you should not interpret the main effects without considering the interaction effect.*

### *1.6 Mention the business implications of performing ANOVA for this particular case study.*

ANOVA allows for the comparison of the relief based on the two independent variables, which in this case is the two active ingredients, A & B (which varies at three different levels).

By observing the comparison between the dependent and independent variables which is 'Relief' and 'A', 'B' respectively in this case, the researchers can vary the levels of 'A' &/or 'B' whichever best treats the hay fever and produce the effective drug in the market.

*Note:*

*That particular brand of drug might perform well in terms of generating larger number of customers, and ultimately increasing the market share.*

***Problem 2:***

A dataset contains the names of various colleges and various parameters of various institutions.

***About the Data:*** This dataset contains 18 variables which are as follows:

1) Names: Names of various university and colleges
2) Apps: Number of applications received
3) Accept: Number of applications accepted
4) Enroll: Number of new students enrolled
5) Top10perc: Percentage of new students from top 10% of Higher Secondary class
6) Top25perc: Percentage of new students from top 25% of Higher Secondary class
7) F.Undergrad: Number of full-time undergraduate students
8) P.Undergrad: Number of part-time undergraduate students
9) Outstate: Number of students for whom the particular college or university is Out-of-state tuition
10) Room.Board: Cost of Room and board
11) Books: Estimated book costs for a student
12) Personal: Estimated personal spending for a student
13) PhD: Percentage of faculties with Ph.D.'s
14) Terminal: Percentage of faculties with terminal degree
15) S.F.Ratio: Student/faculty ratio
16) perc.alumni: Percentage of alumni who donate
17) Expend: The Instructional expenditure per student
18) Grad.Rate: Graduation rate

Basis this data, we are required to perform PCA (Principal Component Analysis).

| Names | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Abilene Christian University | 1660 | 1232 | 721 | 23 | 52 | 2885 | 537 | 7440 | 3300 | 450 | 2200 | 70 | 78 | 18.1 | 12 | 7041 | 60 |
| Adelphi University | 2186 | 1924 | 512 | 16 | 29 | 2683 | 1227 | 12280 | 6450 | 750 | 1500 | 29 | 30 | 12.2 | 16 | 10527 | 56 |
| Adrian College | 1428 | 1097 | 336 | 22 | 50 | 1036 | 99 | 11250 | 3750 | 400 | 1165 | 53 | 66 | 12.9 | 30 | 8735 | 54 |
| Agnes Scott College | 417 | 349 | 137 | 60 | 89 | 510 | 63 | 12960 | 5450 | 450 | 875 | 92 | 97 | 7.7 | 37 | 19016 | 59 |

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

The dimension of the data is: (777, 18) – Without dropping the 'Names' column
(777, 17) – By dropping 'Names' column

There are no missing values in the dataset.

***<u>UNIVARIATE ANALYSIS</u>***

| Variables | Mean | Median | IQR | Standard Deviation | Skewness |
|---|---|---|---|---|---|
| Apps | 3001.64 | 1558 | 2848 | 3870.20 | 3.72 |
| Accept | 2018.80 | 1110 | 1820 | 2451.11 | 3.42 |
| Enroll | 779.97 | 434 | 660 | 929.18 | 2.69 |
| Top10perc | 27.56 | 23 | 20 | 17.64 | 1.41 |
| Top25perc | 55.80 | 54 | 28 | 19.80 | 0.26 |
| F.Undergrad | 3699.91 | 1707 | 3013 | 4850.42 | 2.61 |
| P.Undergrad | 855.30 | 353 | 872 | 1522.43 | 5.69 |
| Outstate | 10440.67 | 9990 | 5605 | 4023.02 | 0.51 |
| Room.Board | 4357.53 | 4200 | 1453 | 1096.70 | 0.48 |
| Books | 549.38 | 500 | 130 | 165.11 | 3.49 |
| Personal | 1340.64 | 1200 | 850 | 677.07 | 1.74 |
| PhD | 72.66 | 75 | 23 | 16.33 | -0.77 |
| Terminal | 79.70 | 82 | 21 | 14.72 | -0.82 |
| S.F.Ratio | 14.09 | 14 | 5 | 3.96 | 0.67 |
| perc.alumni | 22.74 | 21 | 18 | 12.39 | 0.61 |
| Expend | 9660.17 | 8377 | 4079 | 5221.77 | 3.46 |
| Grad.Rate | 65.46 | 65 | 25 | 17.18 | -0.11 |

a. Mean- This is highly affected by outliers; As observed, the average for 'Outstate' is the highest and 'S.F.Ratio' is the lowest.

b. Standard Deviation- It shows us how the data points are spread across the mean. Smaller the value, closer is the data point to mean and vice versa. 'Expend' seems to have the highest standard deviation and 'S.F.Ratio' have the lowest standard deviation.

c. Median- Shows us the $50^{th}$ percentile of the dataset and is the least affected by the outlier.

Here, for variables: 'Apps', 'Accept', 'Enroll', 'Top10perc', 'Top25perc', 'F.Undergrad', 'P.Undergrad', 'Outstate', 'Room.Board', 'Books', 'Personal', 'S.F.Ratio', 'Perc.alumni' and 'Expend' mean is greater than the median implying that these variables seems to be right skewed.

On the other hand, 'PhD' and 'Terminal' have mean less than the median implying that these variables seems to be left skewed.

*Note:*

*One interesting thing that we observe from the above table is even though the Mean of 'Grad.Rate' is higher than its Median, we see that the skewness is negative. This indicates 'Grad.Rate' is also left skewed.*

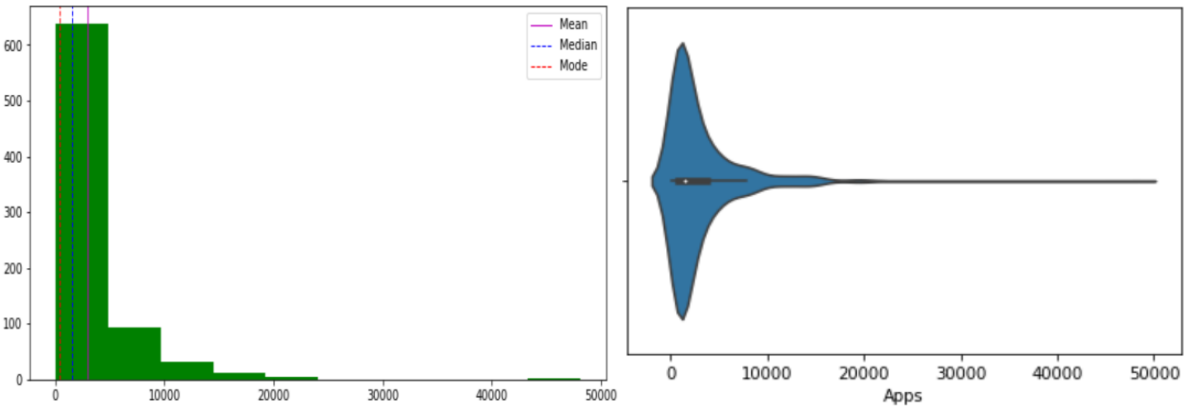
d. Inter Quartile Range- This also helps assess the spread of the data. Higher the IQR, greater is the spread of the data.

Here, for the variables, we see that IQR (the difference between Q1 and Q3) varies to a greater extent. The variable 'Outstate' seems to have the greater spread and 'S.F.Ratio' has the lowest spread.

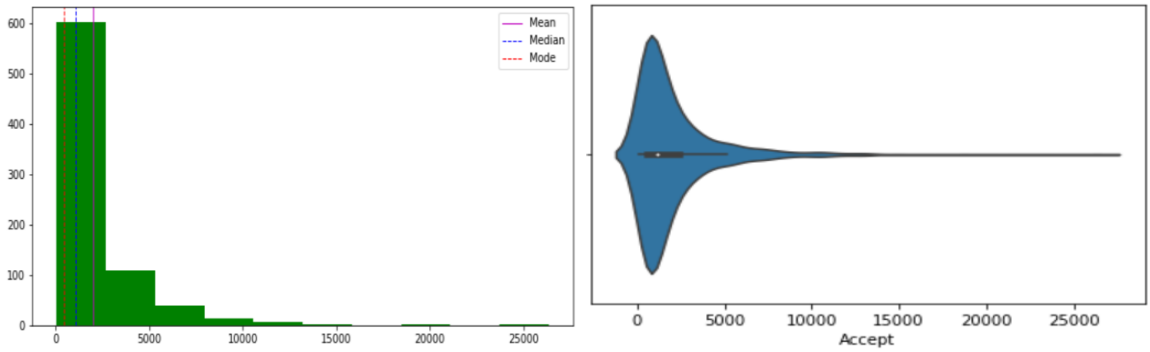
e. Skewness- In layman's language it is the distortion in the normal distribution (the bell-curve) in the dataset.

Though, we can conclude whether the data is skewed or not just by observing Mean and Median of the variables under consideration, Skewness really helps in cases such as 'Grad.Rate' where Mean & Median vary only by 0.46, under such cases we might think the data is right skewed but 'Skewness' tells us exactly how the data is skewed.

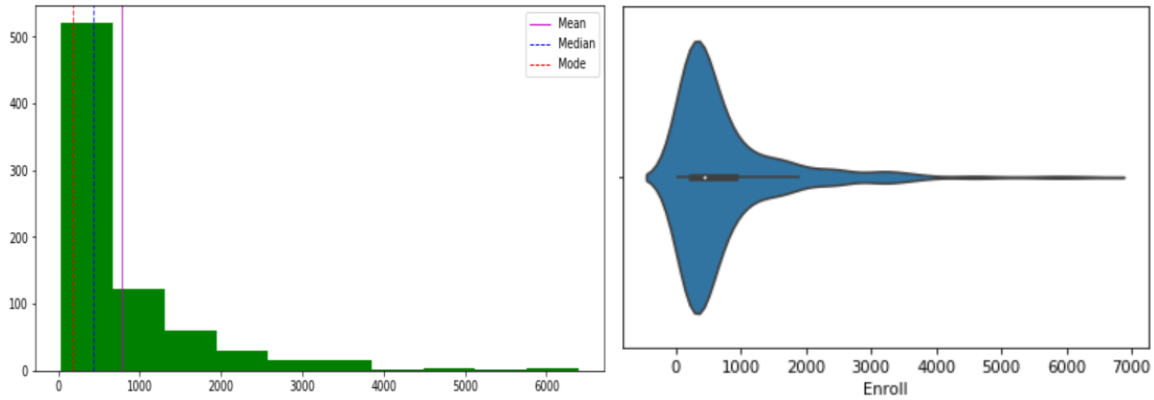
1. Apps



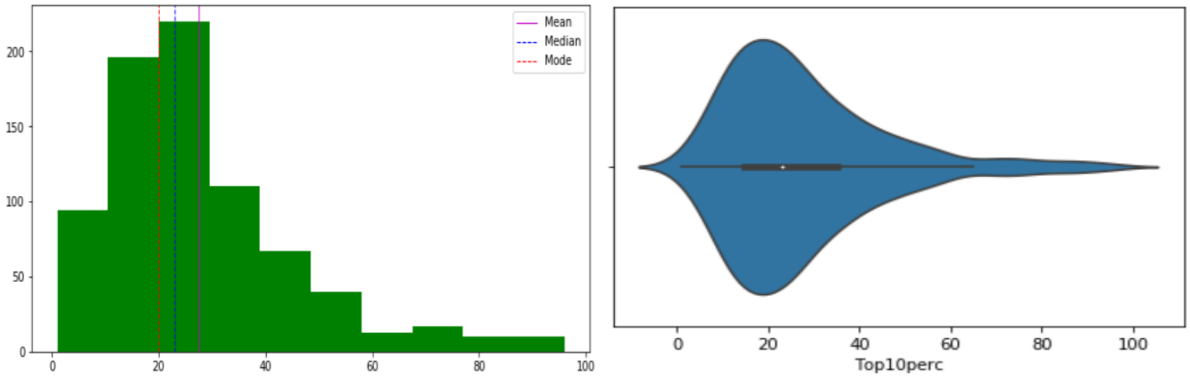

2. Accept



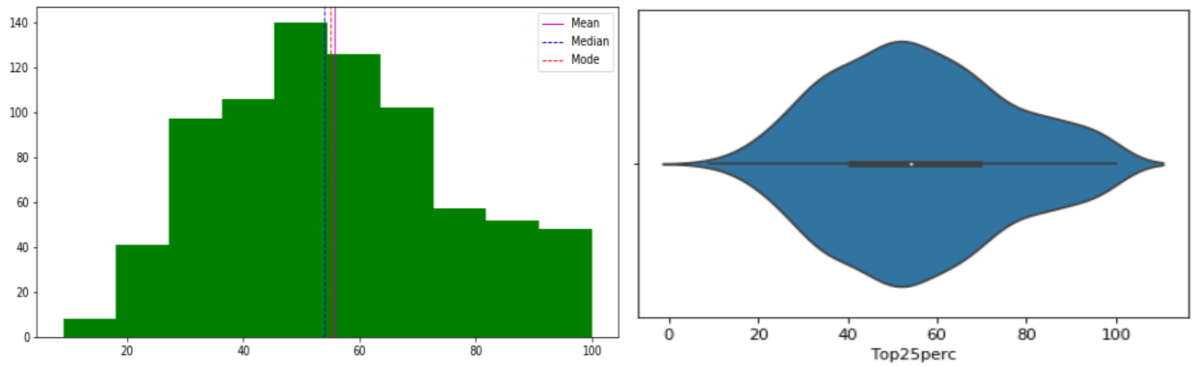
3. Enroll



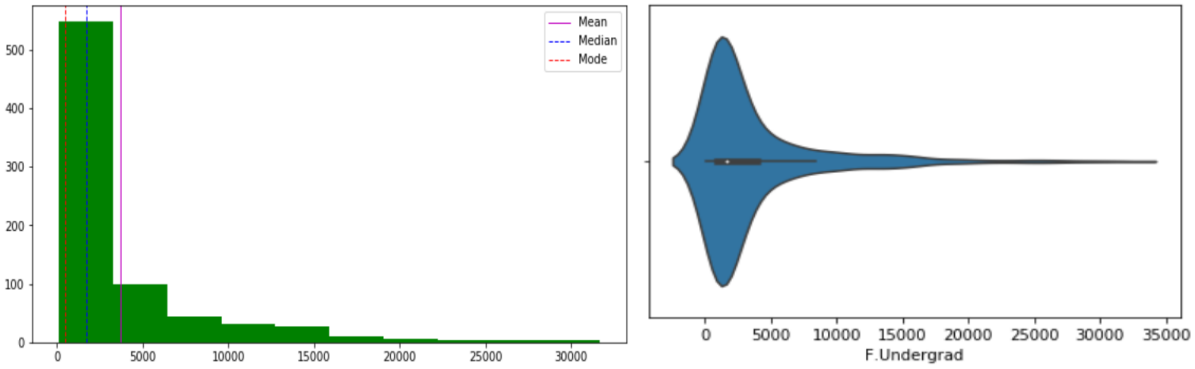

4. Top10perc



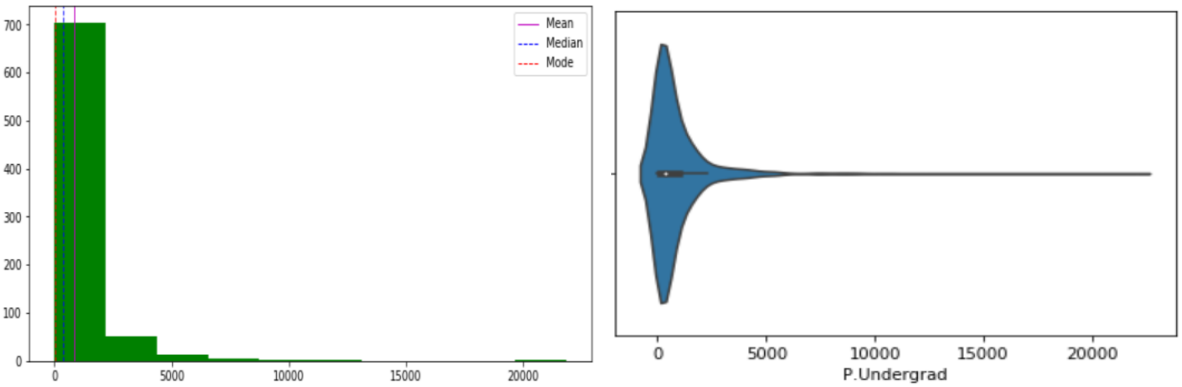

5. Top25perc



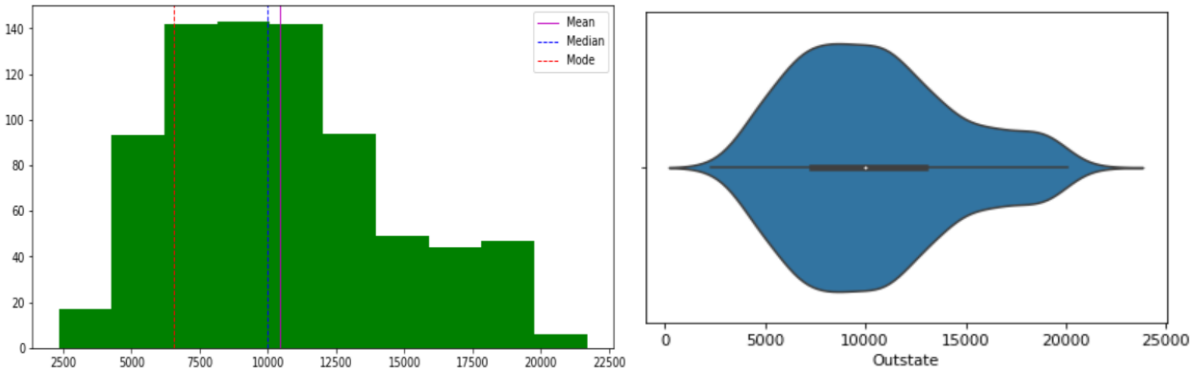
6. F.Undergrad



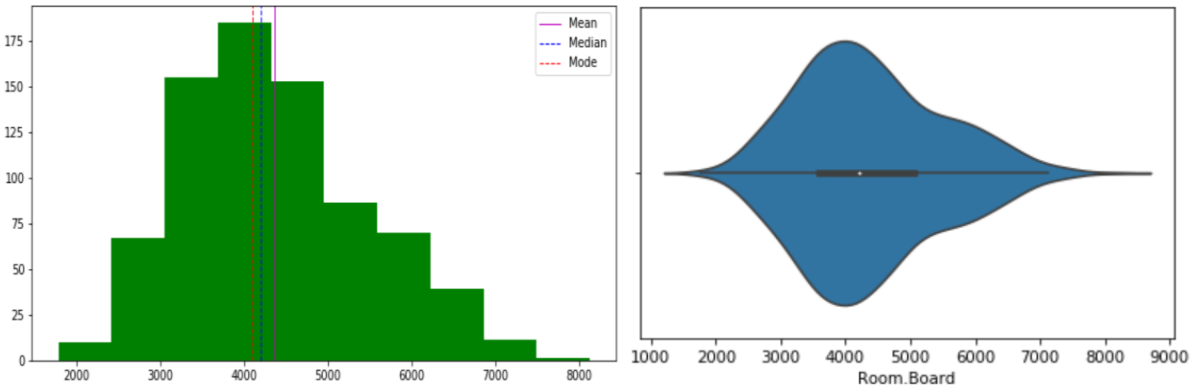

7. P.Undergrad



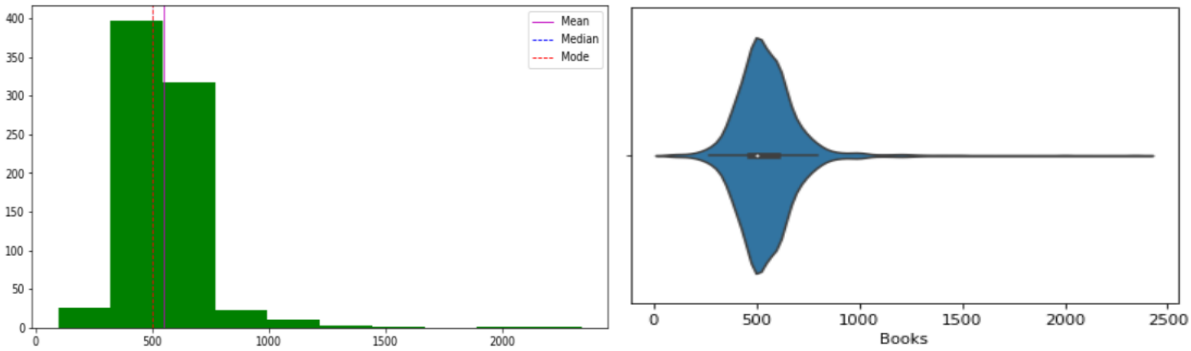

8. Outstate



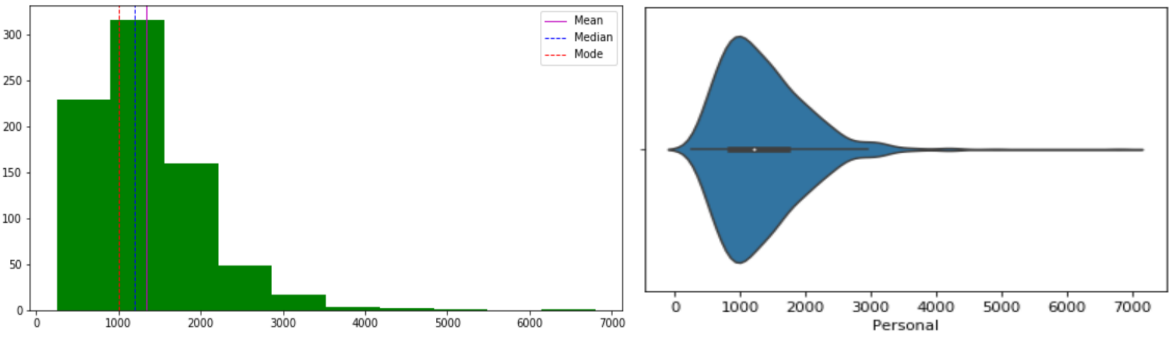
9. Room.Board



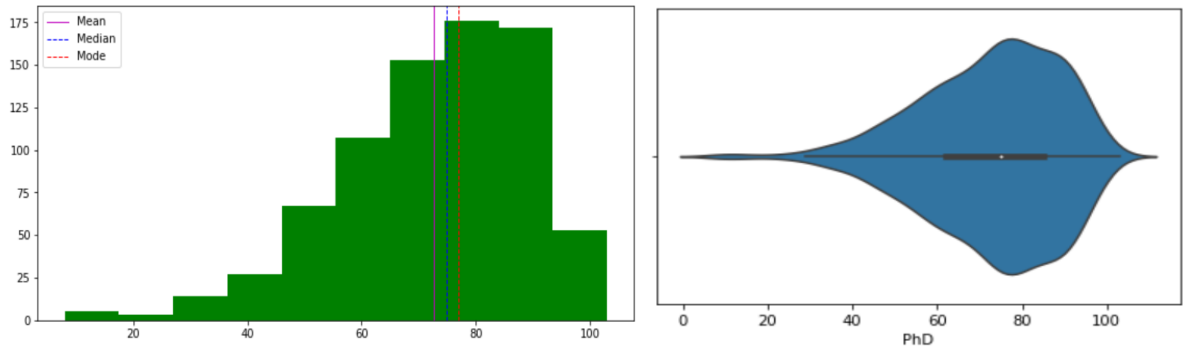

10. Books



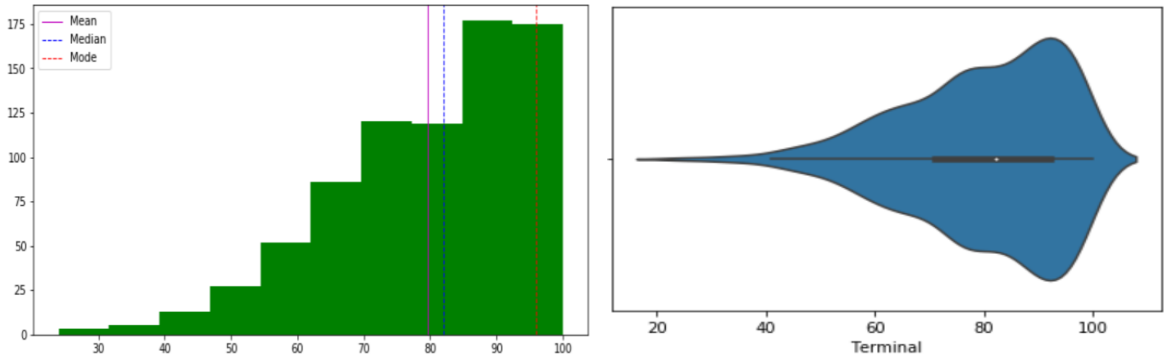

11. Personal



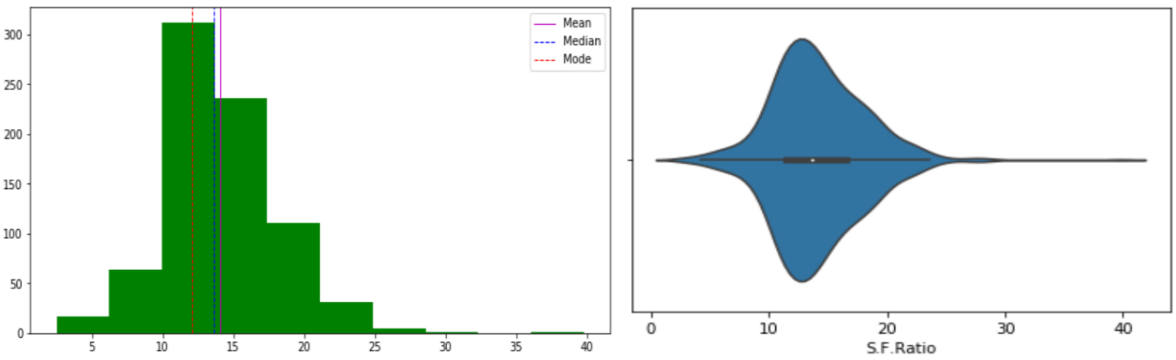
12. PhD



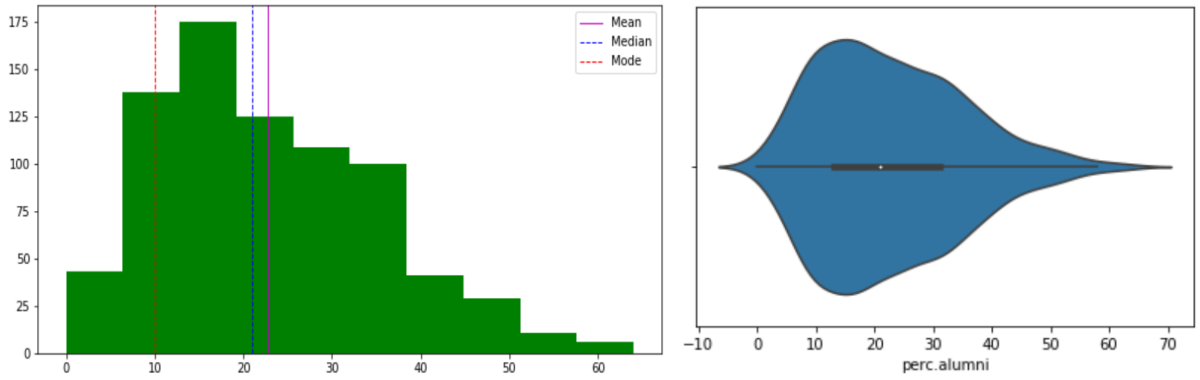

13. Terminal



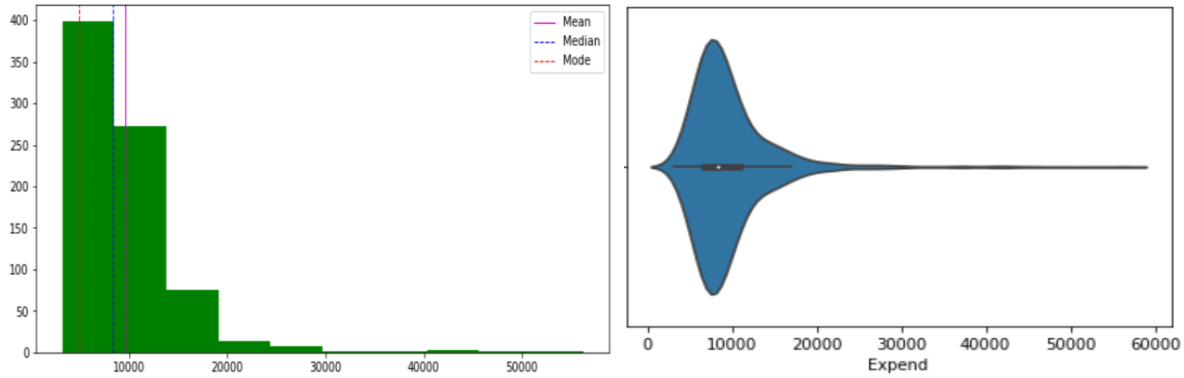

14. S.F.Ratio



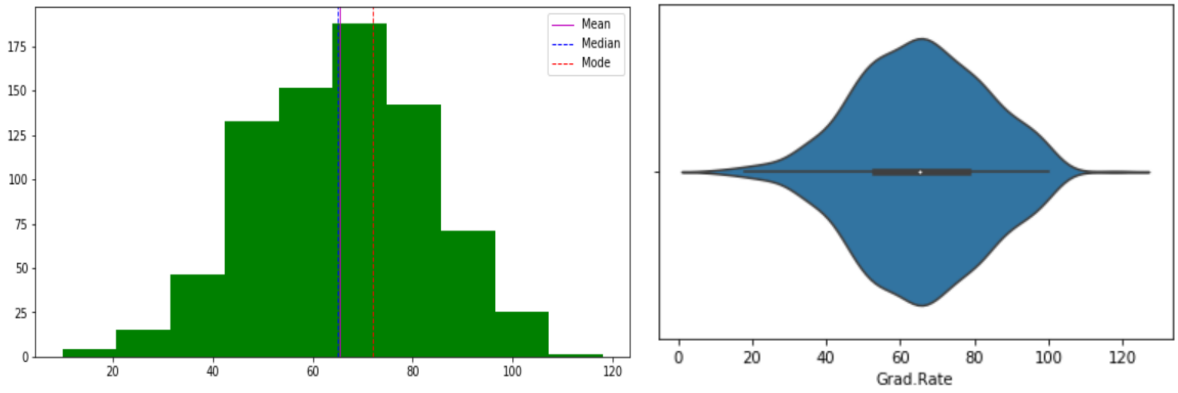
15. Perc.alumni




16. Expend




17. Grad.Rate




As seen from the above Histogram and Violin Plot, most of the variables are right-skewed (implying mean is greater than the median) and only three variables of these: PhD, Terminal and Grad.Rate are left-skewed (as observed from Skewness).

## ***BIVARIATE ANALYSIS***

### I. Covariance

| Column1 | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Apps | 14978460 | 8949860 | 3045256 | 23133 | 26953 | 15289700 | 2346620 | 780970 | 700073 | 84704 | 468347 | 24689 | 21053 | 1465 | -4327 | 5246171 | 9756 |
| Accept | 8949860 | 6007960 | 2076268 | 8321 | 12013 | 10393580 | 1646670 | -253962 | 244347 | 45943 | 333557 | 14238 | 12182 | 1710 | -4859 | 1596272 | 2834 |
| Enroll | 3045256 | 2076268 | 863368 | 2972 | 4173 | 4347530 | 725791 | -581189 | -40997 | 17291 | 176738 | 5029 | 4217 | 873 | -2082 | 311345 | -357 |
| Top10perc | 23133 | 8321 | 2972 | 311 | 312 | 12089 | -2829 | 39907 | 7187 | 346 | -1115 | 153 | 128 | -27 | 100 | 60879 | 150 |
| Top25perc | 26953 | 12013 | 4173 | 312 | 392 | 19159 | -1615 | 38992 | 7200 | 378 | -1084 | 177 | 153 | -23 | 103 | 54546 | 162 |
| F.Undergrad | 15289700 | 10393580 | 4347530 | 12089 | 19159 | 23526580 | 4212910 | -4209843 | -366458 | 92536 | 1041709 | 25212 | 21424 | 5370 | -13792 | 472404 | -6563 |
| P.Undergrad | 2346620 | 1646670 | 725791 | -2829 | -1615 | 4212910 | 2317799 | -1552704 | -102392 | 20410 | 329732 | 3707 | 3181 | 1401 | -5297 | -664351 | -6721 |
| Outstate | 780970 | -253962 | -581189 | 39907 | 38992 | -4209843 | -1552704 | 16184660 | 2886597 | 25808 | -814674 | 25158 | 24164 | -8835 | 28230 | 14133240 | 39480 |
| Room.Board | 700073 | 244347 | -40997 | 7187 | 7200 | -366458 | -102392 | 2886597 | 1202743 | 23170 | -148084 | 5895 | 6047 | -1574 | 3701 | 2873308 | 8005 |
| Books | 84704 | 45943 | 17291 | 346 | 378 | 92536 | 20410 | 25808 | 23170 | 27260 | 20043 | 73 | 243 | -21 | -82 | 96913 | 3 |
| Personal | 468347 | 333557 | 176738 | -1115 | -1084 | 1041709 | 329732 | -814674 | -148084 | 20043 | 458426 | -121 | -305 | 365 | -2399 | -346098 | -3133 |
| PhD | 24689 | 14238 | 5029 | 153 | 177 | 25212 | 3707 | 25158 | 5895 | 73 | -121 | 267 | 204 | -8 | 50 | 36898 | 86 |
| Terminal | 21053 | 12182 | 4217 | 128 | 153 | 21424 | 3181 | 24164 | 6047 | 243 | -305 | 204 | 217 | -9 | 49 | 33733 | 73 |
| S.F.Ratio | 1465 | 1710 | 873 | -27 | -23 | 5370 | 1401 | -8835 | -1574 | -21 | 365 | -8 | -9 | 16 | -20 | -12068 | -21 |
| perc.alumni | -4327 | -4859 | -2082 | 100 | 103 | -13792 | -5297 | 28230 | 3701 | -82 | -2399 | 50 | 49 | -20 | 154 | 27029 | 104 |
| Expend | 5246171 | 1596272 | 311345 | 60879 | 54546 | 472404 | -664351 | 14133240 | 2873308 | 96913 | -346098 | 36898 | 33733 | -12068 | 27029 | 27266870 | 35013 |
| Grad.Rate | 9756 | 2834 | -357 | 150 | 162 | -6563 | -6721 | 39480 | 8005 | 3 | -3133 | 86 | 73 | -21 | 104 | 35013 | 295 |

Covariance shows us the direction between the two variables under consideration. With an increase in one variable if another increase, then it is said to be positive correlation.

It is calculated as follows:

For sample-

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

For population-

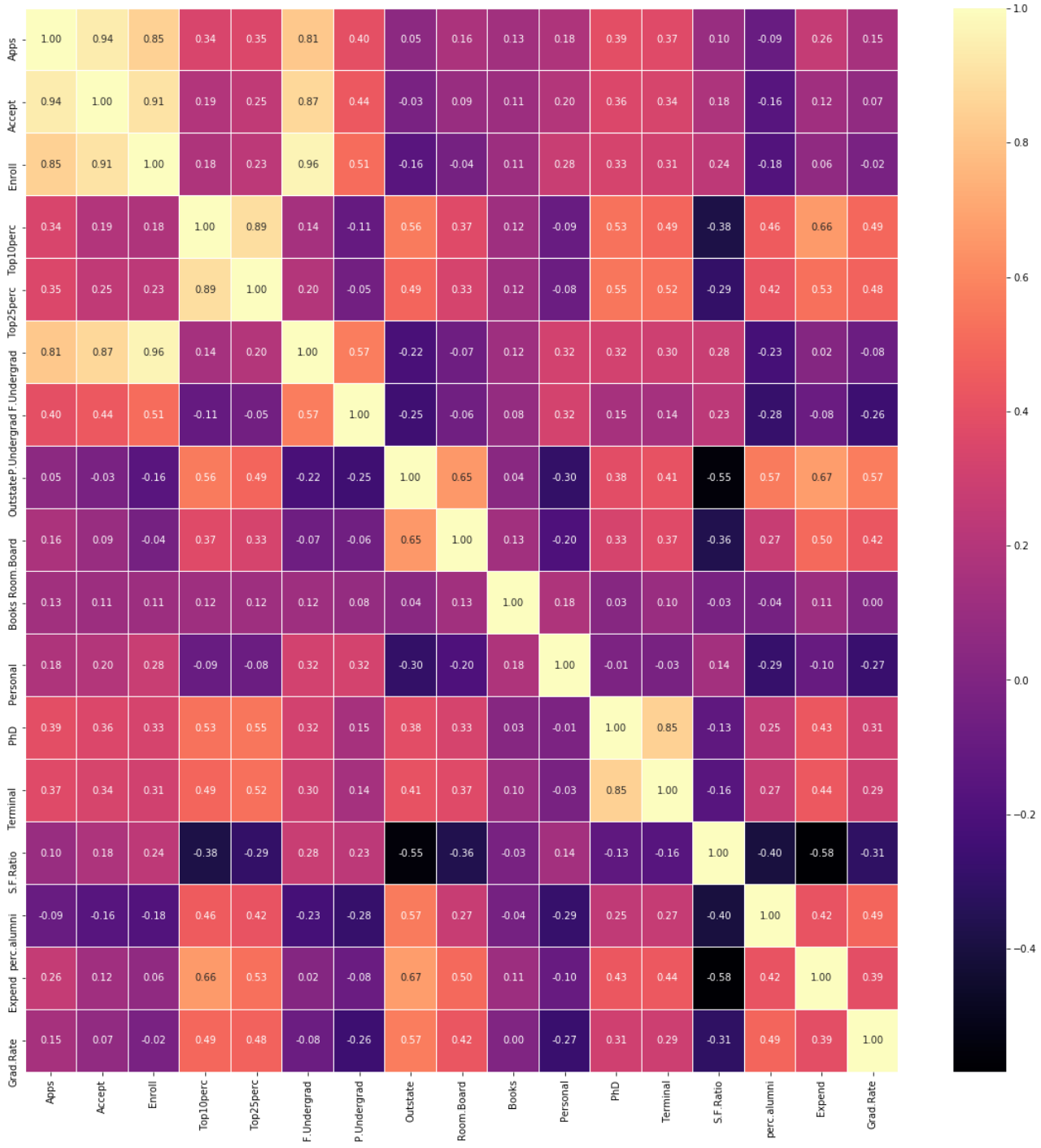
$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Also,

- Cov(x,y) = Cov(y,x)
- Ranges between $-\infty$ and $+\infty$

From the above table, the cells that are highlighted in yellow have negative correlation implying as one variable increases, the other falls (inverse relation), say, a decrease/ increase in 'perc.alumni' will lead to increase/decrease in 'Apps' or vice versa (showing us the direction between these two variables).

## II. Correlation




Correlation shows us the direction as well as the strength of the two variables under consideration.

It is calculated as follows:

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

Also, it ranges from 0 to 1.

Here, we can see that correlation gives the direction as well the strength as compared to covariance which only tells us about the direction of the two variables under consideration.

In this dataset, highest correlation is observed between variables like (Apps, Accept); (Accept, Enroll); (Enroll, F.Undergrad) to name a few- whose correlation is above 0.9, considered strongest.

Lowest correlation is observed between variables like (Accept, Grad.Rate); (Expend, F.Undergrad); (Books, Grad.Rate) etc., whose correlation is below 0.1.

There are also negative correlation between variables like (Personal, Top10perc); (Personal, Top25perc); (perc.alumni, Apps) etc., - Weak negative correlation. (Ranging between -0.1 & 0).

On the other hand, (Expend, S.F.Ratio); (Outstate, S.F.Ratio) have a strong negative correlation (above -0.5).

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

Scaling is a method that converts variables with different scales of measurement into a single scale. This is done in the dataset for easy comparison for the variables under consideration.

Methods of Scaling:

a. Min-Max Scalar
b. Standard Scalar
c. Log Transformation
d. Exponential Transformation

These methods are used according to the different types of data.

Standard scaler is the most commonly used scaling technique. It generally assumes that the given data is distributed normally and will scale them such that the distribution is now centred around (or mean) 0 and standard deviation of 1.

In other words, it removes mean and scales the given data to unit variance. According to an article, "it is a common requirement for many machine learning estimators: they might behave badly if the individual feature do not more or less look like standard normally distributed data...".**

Hence for this dataset, we would proceed with the Standard Scaler.

** [https://oprea.rocks/blog/why-use-sklearn-preprocessing-standardscaler/]

***2.3 <u>Comment on the comparison between covariance and the correlation matrix.</u>***

- The normalized form of covariance is correlation.

- Correlation gives the direction as well the strength whereas covariance only tells us about the direction of the two variables under consideration.

- To measure the relationship between two variables, correlation is preferred over covariance as correlation is unaffected by the change in location and scale.

- Since correlation is the normalized version of covariance, the diagonals of the correlation matrices is always 1.

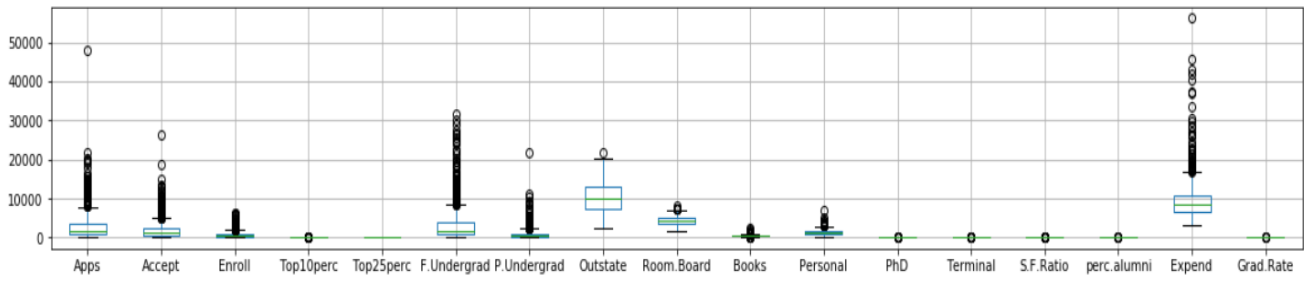
- Also,

$$Cov(x,y) = Cov(y,x) \quad \& \quad Cor(x,y) = Cor(y,x)$$

- After scaling, the covariance and correlation matrix becomes exactly identical.

***2.4 <u>Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.</u>***

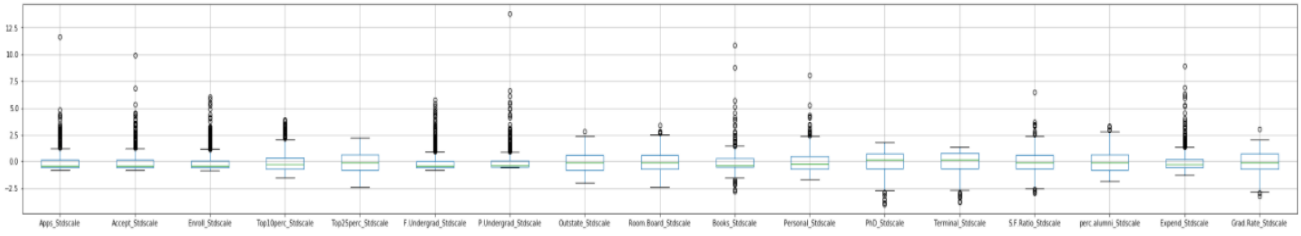
<u>Before Scaling</u>:




**<u>Before Scaling</u>**

After Scaling:

Here, in the 'Education - Post 12th Standard.csv' dataset, we have used Standard Scaler.



**After Scaling**

Prior scaling of this data, the data looked extremely skewed and outliers are extremely spread through the data. On the other hand, after scaling we observe that the data became much more compact in the sense that it is now varying from -3 to +3 and obviously the spread of the outliers are also under a certain range.

### *2.5 Build the covariance matrix and calculate the eigenvalues and the eigenvector.*

## *STANDARD SCALER*

## *Eigen Vectors*

```
%s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
    5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
    9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
    4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
    2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
    5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
    1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
   -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
   -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
   -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
    1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
   -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
    1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
   -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
   -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
   -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
    3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
   -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
   -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
   -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
   -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
```

```
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
[-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
   2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
   4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
   1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
  -8.23443779e-01]
 [-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
   5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
  -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
  -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
   3.54559731e-01]
 [-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
  -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
   1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
   3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
  -2.81593679e-02]
 [ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
  -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
   9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
  -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
  -3.92640266e-02]
 [-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
   1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
   1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
   4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
   2.32224316e-02]
 [-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
   2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
   2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
  -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
   1.64850420e-02]
 [ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
  -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
  -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
   4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
  -1.10262122e-02]
 [-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
  -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
   2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
  -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
   1.82660654e-01]
 [-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
   7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
   4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
   6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
   3.25982295e-01]
 [-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
  -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
  -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
   2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
```

```
   1.22106697e-01]]
```

***Eigen Values***

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

NOTE:

***Covariance Matrix***

```
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304
0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117
0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465
0.87534985
   0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767
0.35621633
   0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373
0.96588274
   0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148
0.33189629
   0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445
0.1414708
  -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665
0.53251337
   0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866
0.19970167
  -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441
0.54656564
   0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167
1.00128866
   0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831
0.3187472
   0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569
0.57124738
   1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384
0.14930637
   0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -
0.21602002
  -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232
0.38347594
   0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -
0.06897917
  -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518
0.32962651
   0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
```

```
 [ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676
0.11569867
   0.08130416  0.03890494  0.12812787  1.00128866  0.17952581
0.0269404
   0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
 [ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441
0.31760831
   0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -
0.01094989
  -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
 [ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564
0.3187472
   0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989
1.00128866
   0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
 [ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506
0.30040557
   0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256
0.85068186
   1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
 [ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852
0.28006379
   0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -
0.13069832
  -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
 [-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -
0.22975792
  -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366
0.24932955
   0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
 [ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713
0.01867565
  -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804
0.43331936
   0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
 [ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -
0.07887464
  -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106
0.30543094
   0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

***2.6 <u>Write the explicit form of the first PC (in terms of Eigen Vectors).</u>***

**PC1 = (0.248766 * Apps) + (0.2076015 * Accept) + (0.176304 * Enroll) + (0.354274 * Top10perc) + (0.344001 * Top25perc) + (0.154641 * F.Undergrad) + (0.026443 * P.Undergrad) + (0.294736 * Outstate) + (0.24903 * Room.Board) + (0.064758 * Books) + (-0.04253 * Personal) + (0.318313 * PhD) + (0.317056 * Terminal) + (-0.17696 * S.F.Ratio) + (0.205082 * perc.alumni) + (0.318909 * Expend) + (0.252316 * Grad.Rate)**

**Where PC1: First PC**

***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.***

## *STANDARD SCALER:*

## *Cumulative Variance Explained*

```
[32.0206282  58.36084263  65.26175919  71.18474841  76.67315352
81.65785448  85.21672597  88.67034731  91.78758099  94.16277251
96.00419883  97.30024023  98.28599436  99.13183669  99.64896227
99.86471628  100.        ]
```

- The Cumulative Variance Explained (%) shows that with an increasing number of principal components or factors, the cumulative variance asymptotically approaches 100%. This gives the percentage of variance accounted for by the first *n* components.

Note:

The cumulative percentage for the second component is the sum of the percentage of variance for the first and second components.

- Three common method for Component selection:

i. Based on Eigenvalue: Here we choose the components with eigenvalues higher than 1.

ii. Cumulative Variance Explained: Here, the chosen factors should explain 70 to 80% of variance at least which is considered as a good amount of variance explained by this method. In this manner, 'Cumulative Variance Explained' helps us deciding the optimum number of principal components.

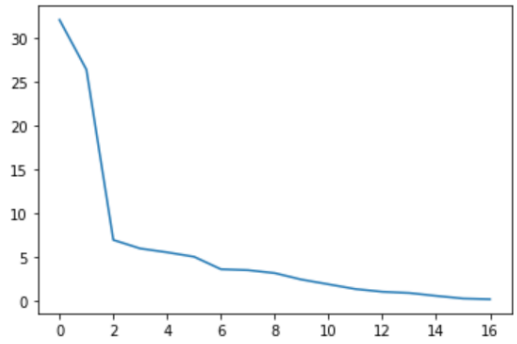
iii. Scree plot: This is a graphical method in which we choose the factors until a break in the elbow.

Hence according to the cumulative variance explained criteria, we in this case choose the 6$^{th}$ component which explains 81.66% of the variance.

- The eigenvectors and eigenvalues of a covariance (or correlation) matrix represent the "core" of a PCA. The eigenvectors (principal components) helps in used to determine the directions of the new feature space. On the other hand, eigenvalues determine their magnitude.

*Note:*

*Also, we can say that eigenvalues explain the variance of the data along the new feature axes.*



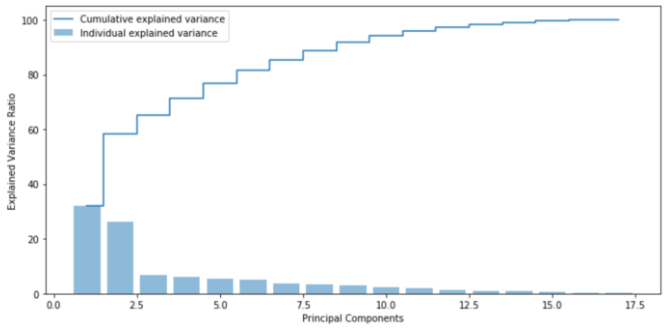

**Scree Plot**

The scree plot is used to determine the optimal number of components. This is a graphical method in which we choose the factors until a break in the elbow. (as discussed above)

The point where the slope of the curve is clearly levelling off ( also known as 'elbow') indicates the optimum number of principal components. In this case, it is '6' principal components.

*Note:*

*From the three common method for Component selection, we can now confirm that the optimum number of principal components is '6'.*




The plot above clearly shows that most of the variance (81.66% of the variance to be precise) can be explained by the first 'six' principal components.

### *2.8 Mention the business implication of using the Principal Component Analysis for this case study.*

- Principal component analysis (PCA) is an efficient solution when we deal with highly correlated variables. These correlated variables are reduced to fewer number of principal component.

- PCA's main objective is to reduce a large set of variables to a smaller set that still retains the originality of the data. In other words, increasing interpretability and minimizing the loss of information.

- PCA is a technique used to emphasize variation and bring out strong patterns in a dataset. It is often used to make data easy to explore and visualize.

Many businesses prefer to reduce the number of variables as they can save money and/or time by removing redundant predictors. Redundant in the sense that those variables might/might not be important but its aspect is covered by other important variable of the dataset, which leads us to eliminating those variables and minimize the information loss and reduce or eradicate the problem of multicollinearity.