ADVANCE STATISTICS PROJECT

Data Analysis Report

Prepared By

JAI GOUTHAM

Submitted on
18-10-2020

Q 1.1

Summary

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. The data on hours of relief can be found in the following .csv file: Fever.csv

Insights:

- From the above problem statement, it indicates us to find out the Compounds varied at three levels at different ingredients A and B separately using One-way anova.
- We have to find out at one of the compounds differs in ingredient A and B.

Approach:

- Using ANOVA for ingredient A and B separately with respect to relief variable.
- We can find the Interaction between the Ingredients A and B with respect to Relief.

For Ingredient A

Step1: State the Null and alternative hypothesis:

H*o*: μ1 = μ2 = μ3 The mean of all Compounds for the A ingredient is same.

H*a*: At least one of the means of all Compounds for the A ingredient is different.

Step2: Decide Level of significance:

Generally, the level of significance will be 5% (α = 0.05)

Step3: Assumptions:

Checking the normality of the data by Two Methods:

3.1) Shapiro Test (Statistical Method)

3.2) Histogram (Visual Method)

Checking the variance of the data by Two Methods:

3.3) Boxplot (Visual Method)

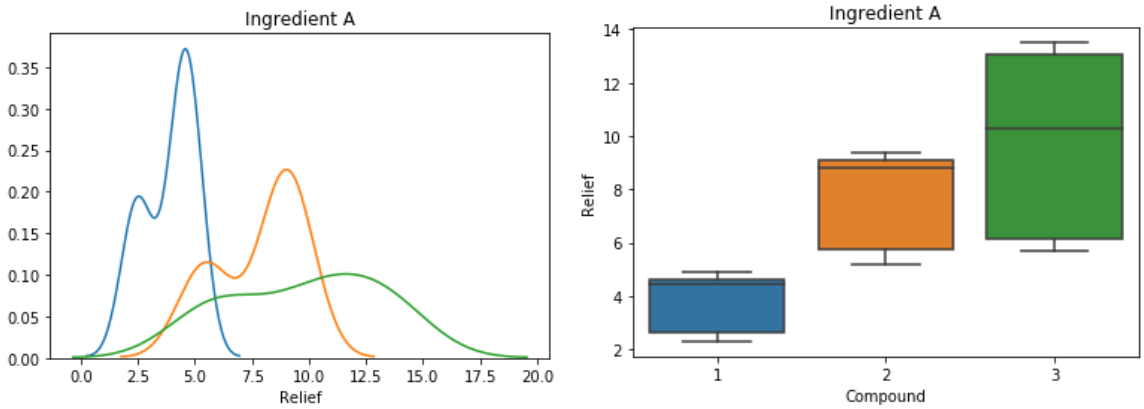
3.4) Levene Test (Statistical Method)

3.1 Shapiro test:

The shapiro test P-value for Compound 1: 0.042, Compound 2: 0.001, Compound 3 is: 0.034

(Note:- Refer python Code File)

3.2, 3.3 Histogram, Boxplot:



3.4 Levene test:

The Levene test P-value for is: 0.0185

(Note:- Refer python Code File)

- Statistical method : From Shapiro test all the samples having P-value is less than α (0.05). Reject Ho(null Hypothesis).
- Visual Method: From the histogram the data looks not normal.
- Statistical method : From Levene test all the samples having P-value is less than α (0.05). Reject Ho (null Hypothesis) .
- Visual Method: From the Box Plot the variance are not equal for the compounds of ingredient A.

Step 4: Inferential Test

1.Parametric test -Method (ANOVA TEST)

2.Non parametric test- Method (Kruskal Wallis test)

- Hence, from the Assumptions of both statistical method (Shapiro test: P-value is less than 5%), (levene test: P-value is less than 5%) and visual method from the Histogram and Box Plot, data looks not normal and having not equal variance, So we are going for non parametric test (Kruskalwallis method to find the Hypothesis).

- The Kruskalwallis test P-value for is: **`2.6992992738200464e-06`**

Step 5: Decision:

Since the p value in this scenario is less than α (0.05), we can say that Reject the Null Hypothesis ($H0$). As a result, we know that at least one of the compound is differs from the other category of the Ingredient A

NOTE: In project as per question 1.2 asking to find one-way anova for variable A with relief

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.02	110.010000	23.465387	4.578242e-07
Residual	33.0	154.71	4.688182	NaN	NaN

- The ANOVA test P-value for is: **`4.578242e-07`**

Since the p value in this scenario is less than α (0.05), we can say that "Reject the Null Hypothesis" (Ho).

As a result, we know that at least one of the compounds is differs from the other category of the Ingredient A.

Conclusion:

- **From both the ANOVA test and kruskalwallis test, the P-value is less than alpha (p-value < α 0.05). So, Reject the Null Hypothesis Ho.**
- **Hence at least one of the means of the compounds differs from other groups in the ingredient A.**

Q 1.3

For Ingredient B

Step1: State the Null and alternative hypothesis:

H*o*: µ1 = µ2 = µ3 The mean of all Compounds for the B ingredient is same.

H*α*: At least one of the means of all Compounds for the B ingredient is different.

Step2: Decide Level of significance:

Generally, the level of significance will be 5% (α = 0.05)

Step3: Assumptions:

Checking the normality of the data by Two Methods:

3.1) Shapiro Test (Statistical Method)

3.2) Histogram (Visual Method)

Checking the variance of the data by Two Methods:

3.3) Boxplot (Visual Method)

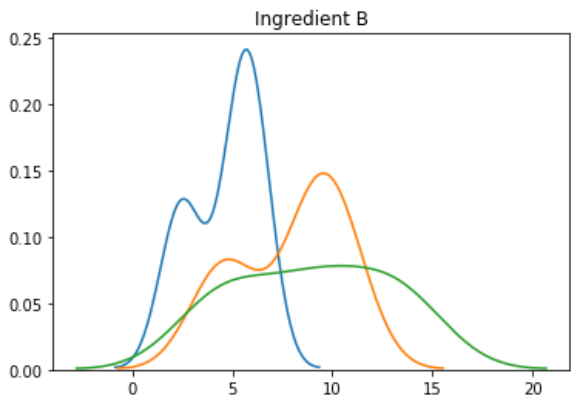
3.4) Levene Test (Statistical Method)

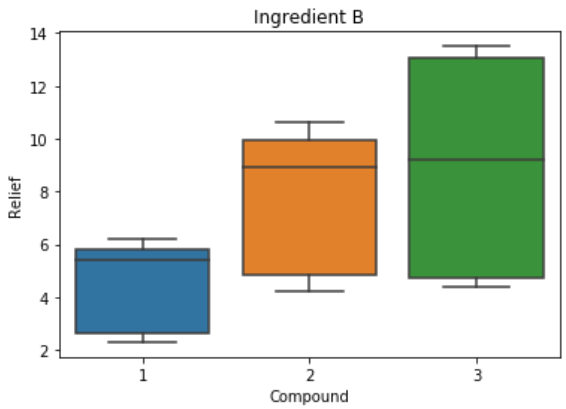
3.1 Shapiro test:

The shapiro test P-value for Compound 1: 0.003, Compound 2: 0.011, Compound 3 is: 0.031

(Note:- Refer python Code File)

3.2, 3.3 Histogram, Boxplot:





3.4 Levene test:

The Levene test P-value for is: 0.0185

(Note:- Refer python Code File)

- Statistical method : From Shapiro test all the samples having P-value is less than α (0.05). Reject Ho(null Hypothesis).
- Visual Method: From the histogram the data looks not normal.
- Statistical method : From Levene test all the samples having P-value is less than α (0.05). Reject Ho (null Hypothesis) .
- Visual Method: From the Box Plot the variance are not equal for the compounds of ingredient A.

Step 4: Inferential Test:

1.Parametric test -Method (ANOVA TEST)

2.Non parametric test- Method (Kruskal Wallis test)

- Hence, from the Assumptions of both statistical method (Shapiro test: P-value is less than 5%), (levene test: P-value is greater than 5%) and visual method from the Histogram and Box Plot, data looks not normal and having slightly equal variance, So we are going for parametric test (Anova Test to find the Hypothesis).

	df	sum_sq	mean_sq	F	PR(>F)
C(B)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

- The ANOVA test P-value for is: 0.00135

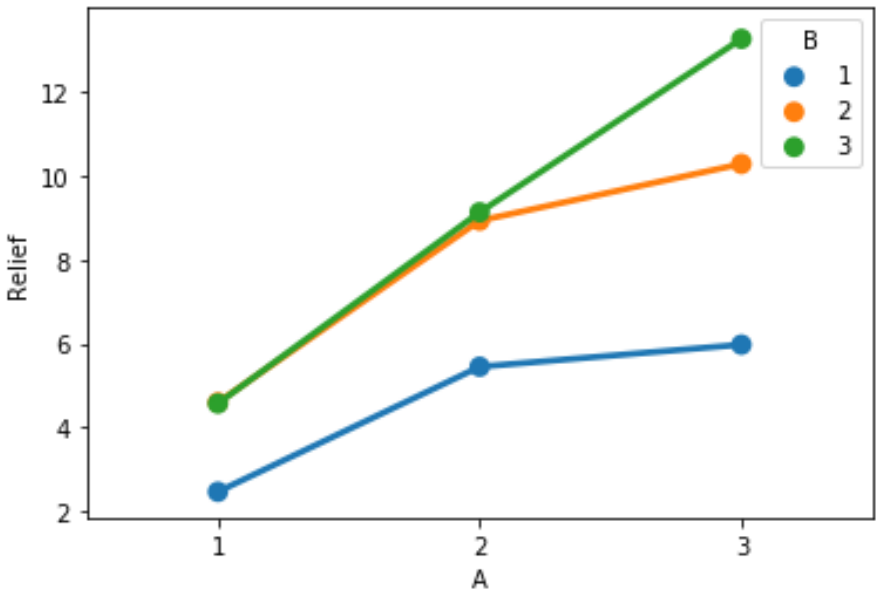
Since the p value in this scenario is less than α (0.05), we can say that “Reject the Null Hypothesis” (Ho).

Conclusion:

- **From both the ANOVA test, the P-value is less than alpha (p-value < α 0.05). So, Reject the Null Hypothesis Ho.**
- **Hence at least one of the means of the compounds differs from other groups in the ingredient B.**

Q 1.4

Checking the Interaction of the compounds with the Ingredients A and B



As seen from the above two interaction plots, **visually there is no interaction** with the two variables of the compounds for Ingredients A and B.

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.020	110.010000	1827.858462	1.514043e-29
C(B)	2.0	123.660	61.830000	1027.329231	3.348751e-26
C(A):C(B)	4.0	29.425	7.356250	122.226923	6.972083e-17
Residual	27.0	1.625	0.060185	NaN	NaN

Conclusion:

- **The P value is very less than alpha**
- **So, From the statistical test (ANOVA Test) there is No interaction with the two variables of the compounds for Ingredients A and B.**

Q 1.5

Performing two way anova for Ingredients A and B

State the Null and alternative hypothesis:

Ho: The mean of Relief variable is equal for both the ingredients A and B is same.

Ha: At least one of the mean of the Relief variable for the ingredient A and B is not same.

Anova test:

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.02	110.010000	109.832850	8.514029e-15
C(B)	2.0	123.66	61.830000	61.730435	1.546749e-11
Residual	31.0	31.05	1.001613	NaN	NaN

Decision:

Considering both the Ingredients (A and B), both the p value is <0.05, So Reject the Null hypothesis

Hence, At least one of the means of the Relief variable for the ingredient A and B is not same.

Q 1.6

Business Implication:

From the Hypothesis , we can decide that there is a difference of average on the relief time on both Ingredient A and B across the three different compounds.

So, When comparing both the ingredients A and B there is no significant Interaction between the variables with respect to the relief time.

The “Hay Fever” relief time does not change any difference in the ingredients, but the mean of the compounds in both ingredients A and B has a difference in it, so with respect to the variable relief the compounds of both ingredients which has variation in the average relief time

Summary

The dataset Education - Post 12th Standard.csv is a dataset that contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx

Insights:

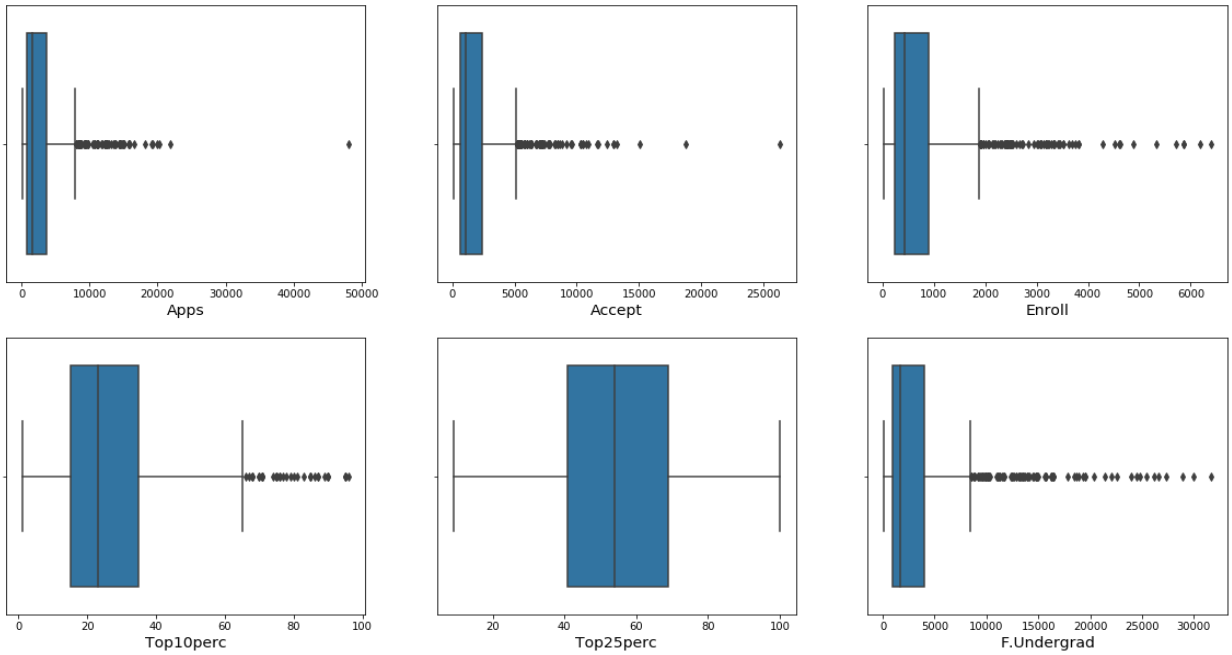
- Principal component analysis is a technique for dimension reduction — so it combines input variables in a specific way, to drop the “least important” variables while still retaining the most valuable parts of all of the variables

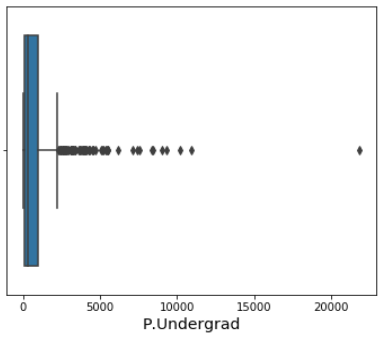
Approach:

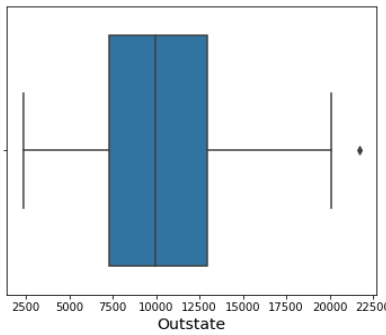
- Using Principal component Analysis the we are going to find the 90% of the variance which is extracted from all variables present in the dataset.

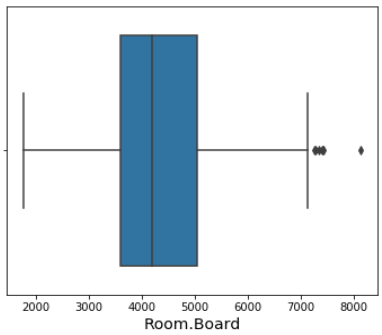
Univariant Analysis:

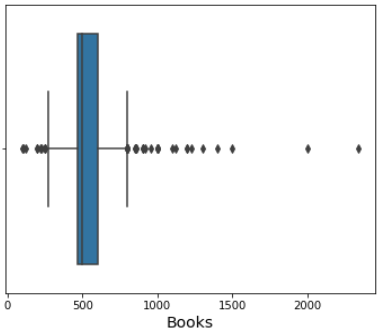
- Checking the Variance using Box plot for all variables in the dataset.

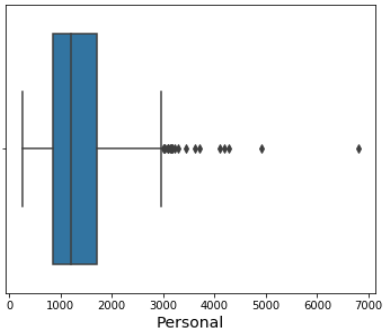


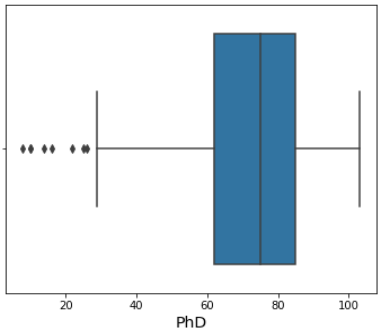


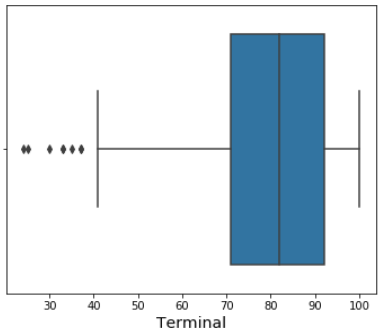


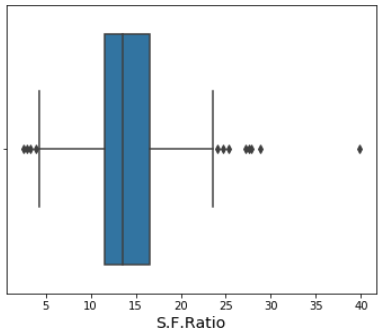


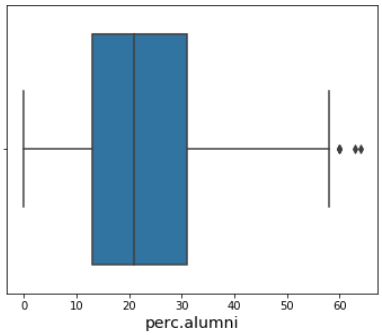


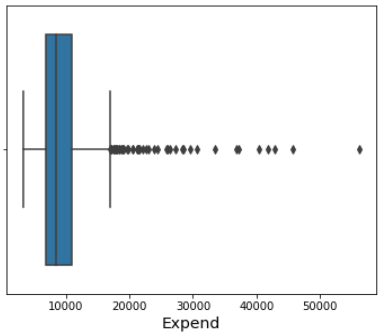


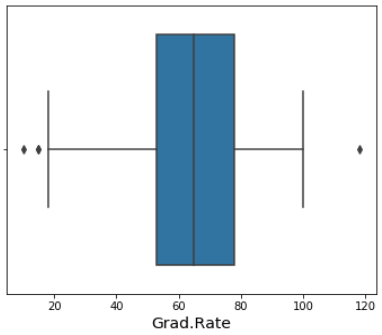




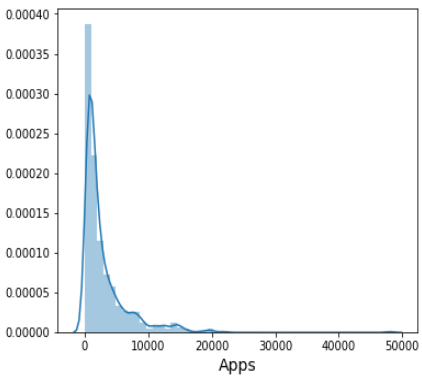


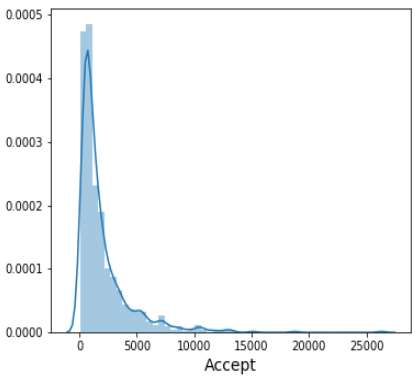


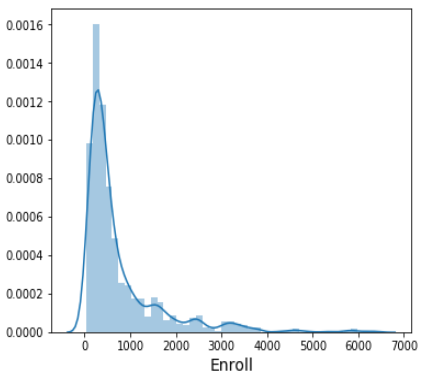


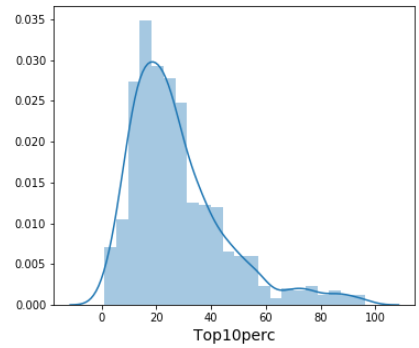


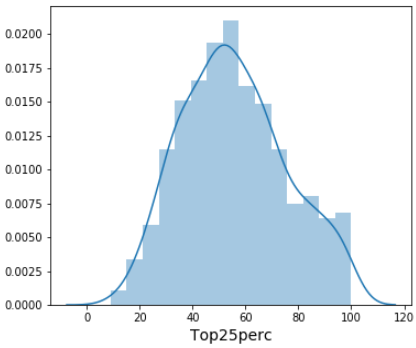
- Checking the Normality using Histogram for all variables in the dataset.

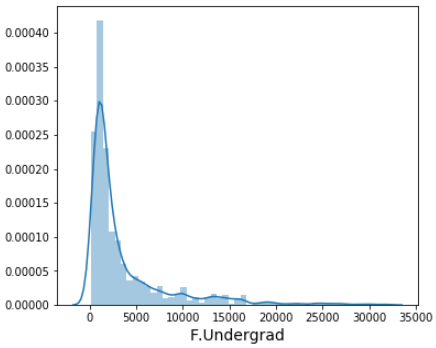


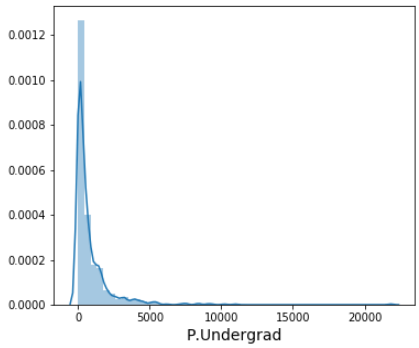


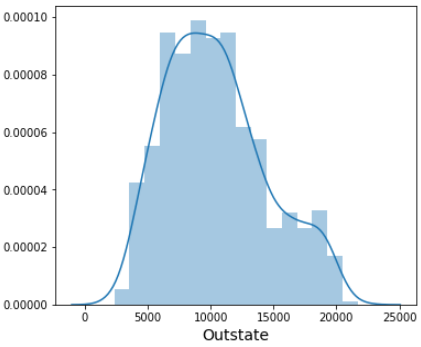


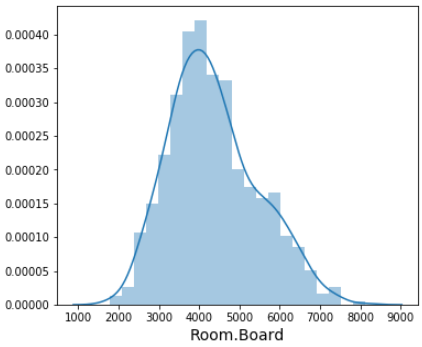

Top10perc

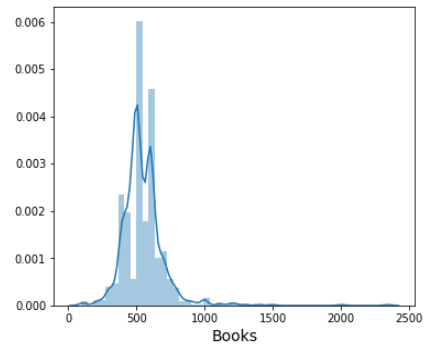

Top25perc

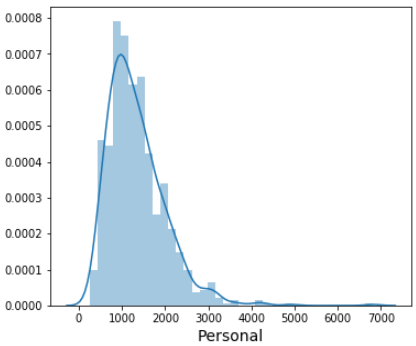

F.Undergrad

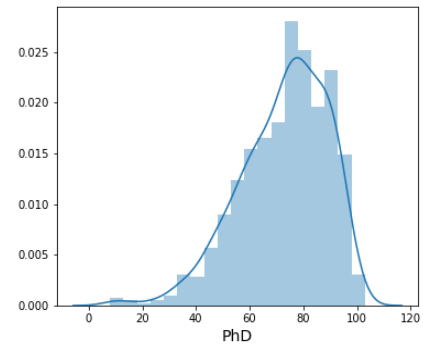

P.Undergrad

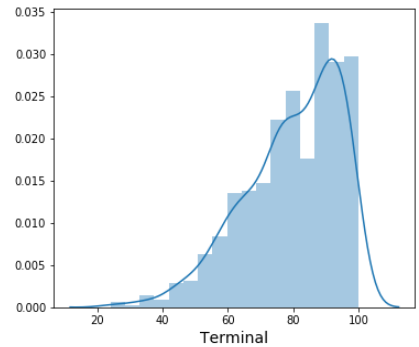

Outstate

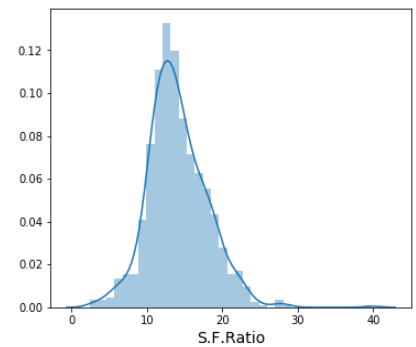

Room.Board

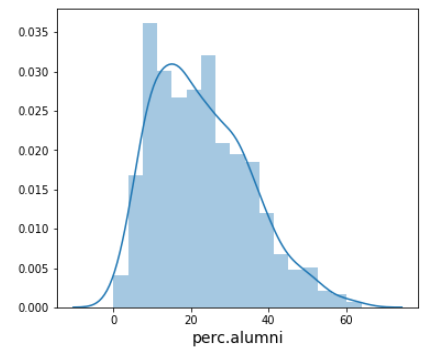

Books

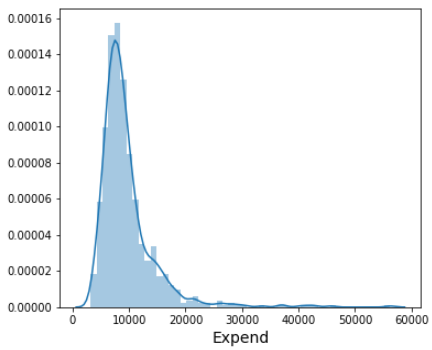

Personal

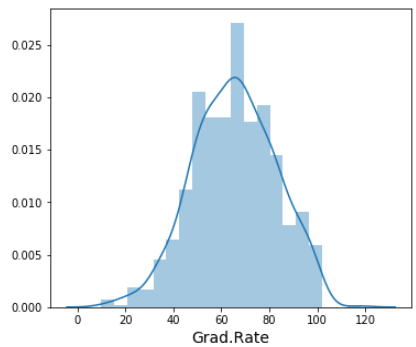

PhD


Terminal


S.F.Ratio

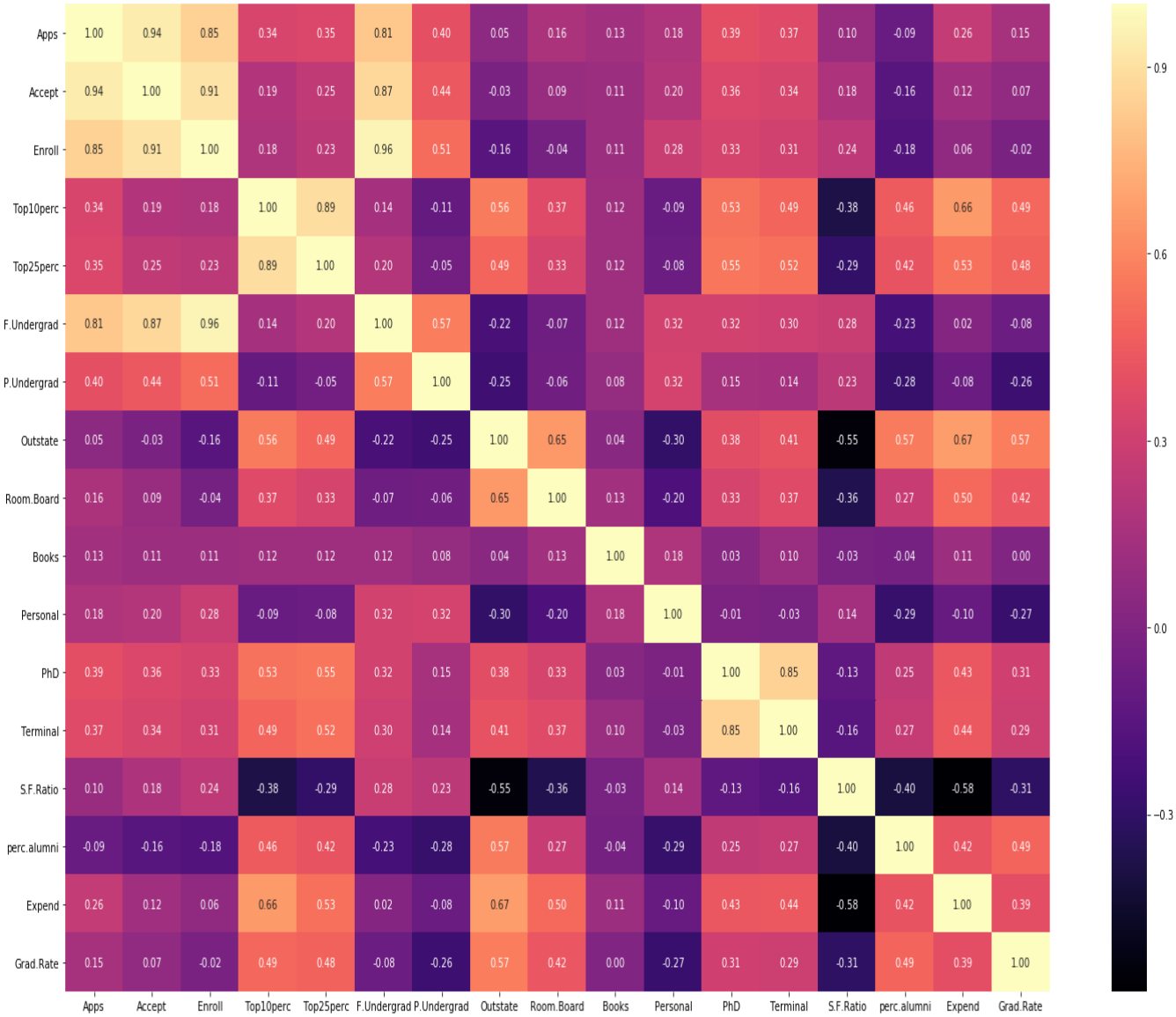

perc.alumni


Expend


Grad.Rate

Multivariant Analysis:

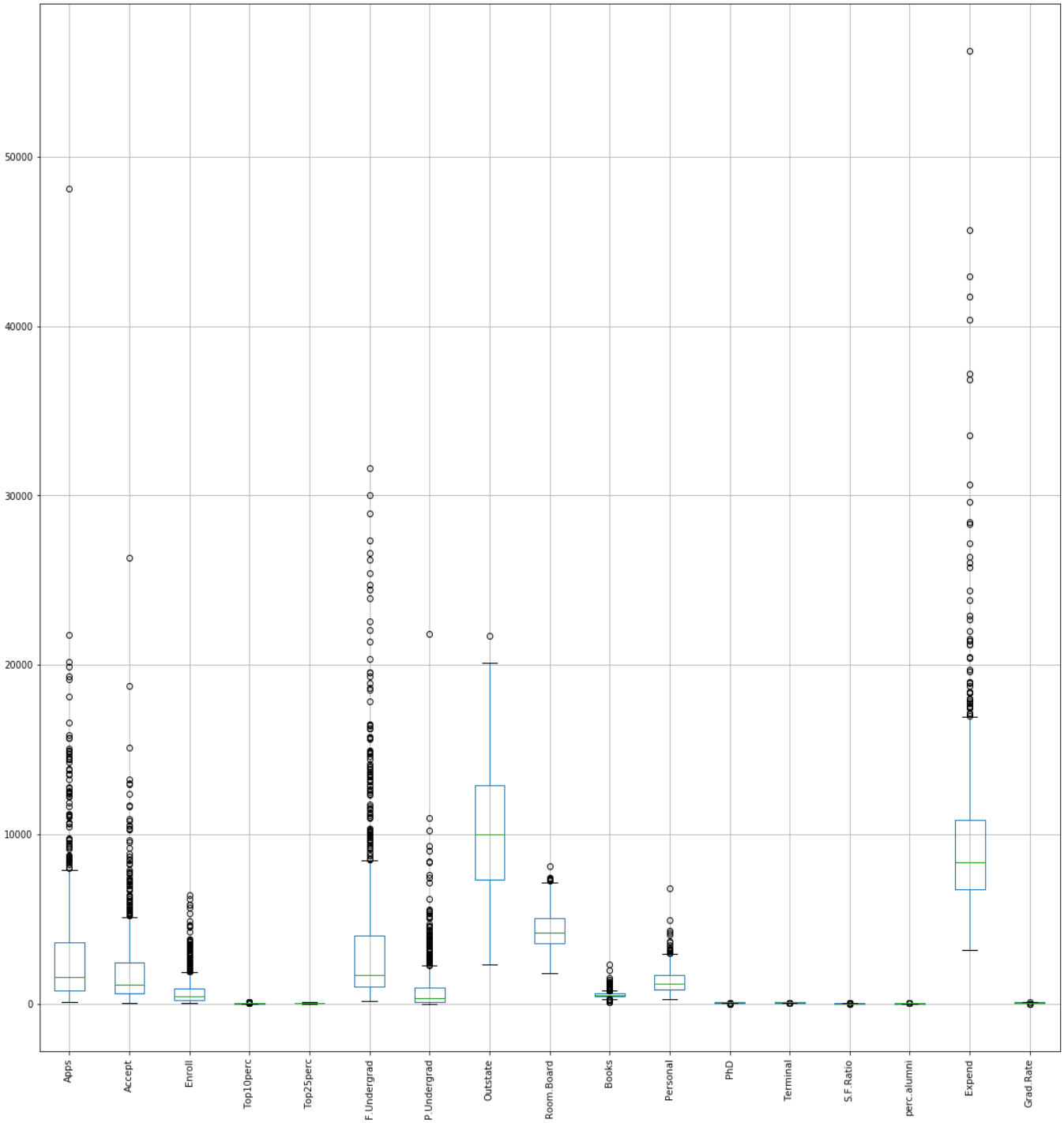
- Multivariate analysis is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time.
- In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related.



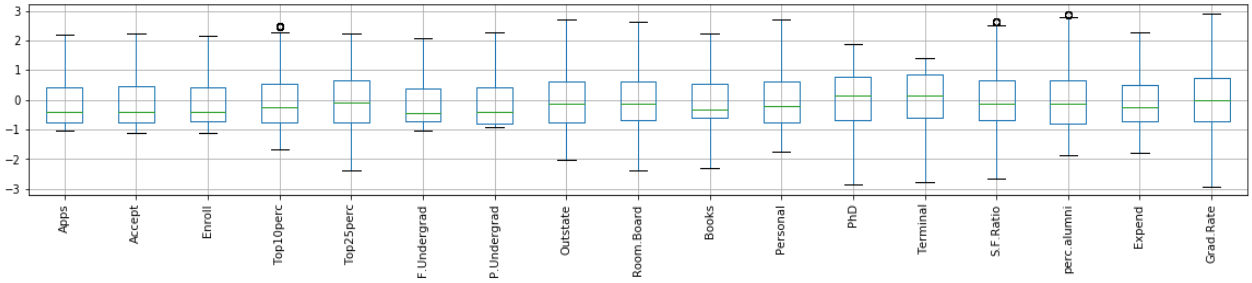
Q 2.4

Outlier check with boxplot before Scaling:

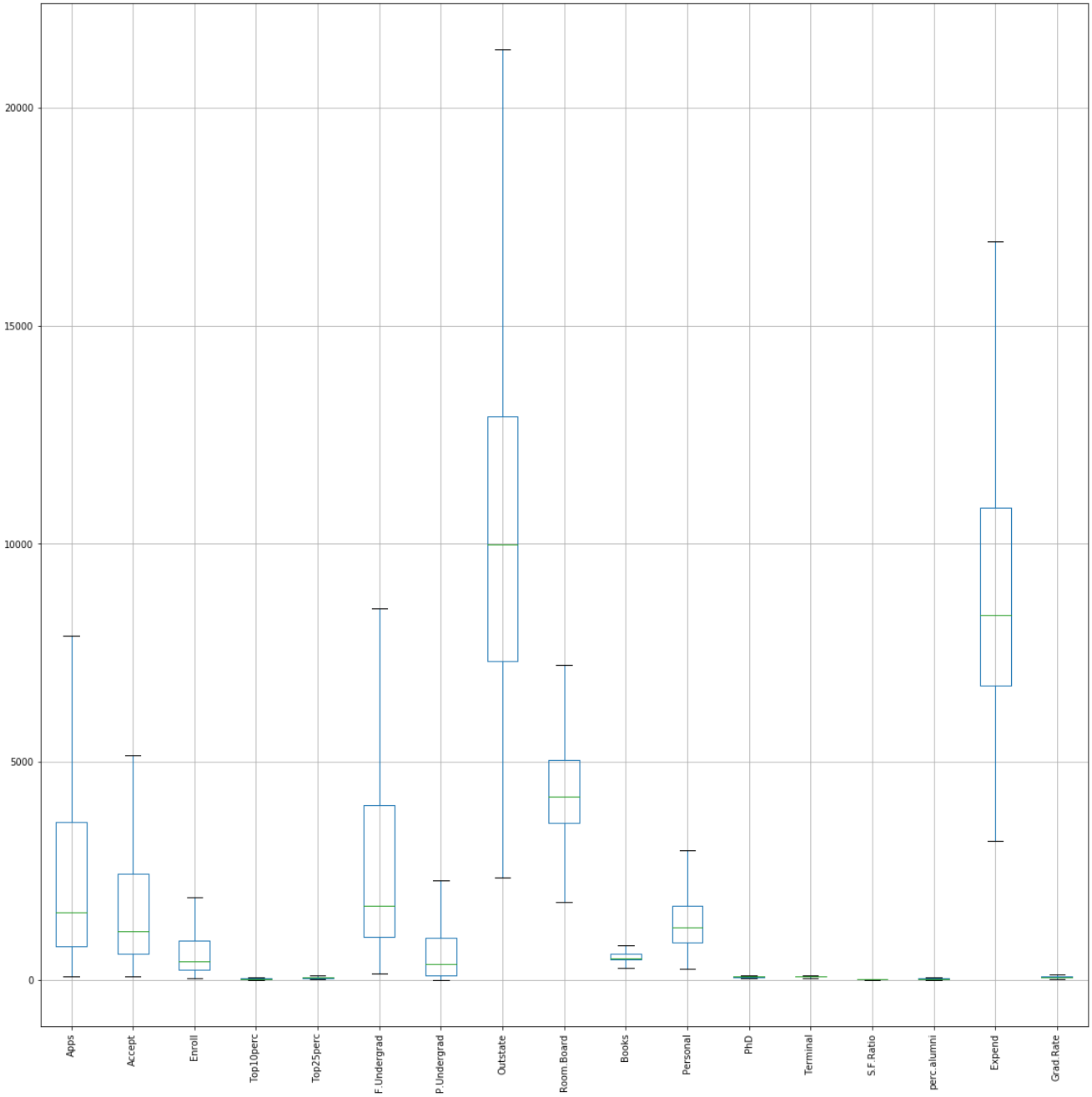
- An outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.



Outlier check with boxplot After Scaling:



Plotting boxplot after outlier treatment:



PRINCIPAL COMPONENT ANALYSIS

PCA is a statistical technique and uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. PCA also is a tool to reduce multidimensional data to lower dimensions while retaining most of the information. Principal Component Analysis (PCA) is a well-established mathematical technique for reducing the dimensionality of data, while keeping as much variation as possible.

Q 2.2

Step1 Standardizing before doing PCA:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

The formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization. So in this dataset we have outliers so we are applying this technique.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.376493	-0.337830	0.106380	-0.246780	-0.191827	-0.018769	-0.166083	-0.746480	-0.968324	-0.776567	1.438500	-0.174045	-0.123239	1.070602	-0.870466	-0.630916	-0.319205
1	-0.159195	0.116744	-0.260441	-0.696290	-1.353911	-0.093626	0.797856	0.457762	1.921680	1.828605	0.289289	-2.745731	-2.785068	-0.489511	-0.545726	0.396097	-0.552693
2	-0.472336	-0.426511	-0.569343	-0.310996	-0.292878	-0.703966	-0.777974	0.201488	-0.555466	-1.210762	-0.260691	-1.240354	-0.952900	-0.304413	0.590864	-0.131845	-0.669437
3	-0.889994	-0.917871	-0.918613	2.129202	1.677612	-0.898889	-0.828267	0.626954	1.004218	-0.776567	-0.736792	1.205884	1.190391	-1.679429	1.159159	2.287940	-0.377577
4	-0.982532	-1.051221	-1.062533	-0.696290	-0.596031	-0.995610	0.297726	-0.716623	-0.216006	2.219381	0.289289	0.202299	-0.538069	-0.568839	-1.682316	0.512468	-2.916759

Q 2.3

Step2 Creating covariance matrix after standardization:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1.001289	0.956538	0.898039	0.321756	0.364961	0.862111	0.520493	0.065421	0.187717	0.236442	0.230244	0.464522	0.435038	0.126574
1	0.956538	1.001289	0.936482	0.223586	0.274033	0.898190	0.573429	-0.005009	0.119740	0.208974	0.256676	0.427891	0.403929	0.188749
2	0.898039	0.936482	1.001289	0.171977	0.230731	0.968549	0.642422	-0.155856	-0.023876	0.202317	0.339785	0.382031	0.354836	0.274622
3	0.321756	0.223586	0.171977	1.001289	0.915053	0.111358	-0.180241	0.562884	0.357826	0.153650	-0.116880	0.544749	0.507401	-0.388426
4	0.364961	0.274033	0.230731	0.915053	1.001289	0.181429	-0.099423	0.490200	0.331413	0.169980	-0.086922	0.552172	0.528334	-0.297616
5	0.862111	0.898190	0.968549	0.111358	0.181429	1.001289	0.697027	-0.226457	-0.054546	0.208147	0.360246	0.362030	0.335486	0.324922
6	0.520493	0.573429	0.642422	-0.180241	-0.099423	0.697027	1.001289	-0.354673	-0.067725	0.122686	0.344496	0.127827	0.122309	0.371085
7	0.065421	-0.005009	-0.155856	0.562884	0.490200	-0.226457	-0.354673	1.001289	0.656334	0.005117	-0.326029	0.391825	0.413110	-0.574422
8	0.187717	0.119740	-0.023876	0.357826	0.331413	-0.054546	-0.067725	0.656334	1.001289	0.109065	-0.219837	0.341909	0.379759	-0.376915
9	0.236442	0.208974	0.202317	0.153650	0.169980	0.208147	0.122686	0.005117	0.109065	1.001289	0.240172	0.136566	0.159523	-0.008547
10	0.230244	0.256676	0.339785	-0.116880	-0.086922	0.360246	0.344496	-0.326029	-0.219837	0.240172	1.001289	-0.011699	-0.032012	0.174137
11	0.464522	0.427891	0.382031	0.544749	0.552172	0.362030	0.127827	0.391825	0.341909	0.136566	-0.011699	1.001289	0.864040	-0.129556
12	0.435038	0.403929	0.354836	0.507401	0.528334	0.335486	0.122309	0.413110	0.379759	0.159523	-0.032012	0.864040	1.001289	-0.151188
13	0.126574	0.188749	0.274622	-0.388426	-0.297616	0.324922	0.371085	-0.574422	-0.376915	-0.008547	0.174137	-0.129556	-0.151188	1.001289
14	-0.101288	-0.165729	-0.223010	0.456384	0.417369	-0.285825	-0.419874	0.566465	0.272744	-0.042887	-0.306147	0.249198	0.266375	-0.412632
15	0.243248	0.162017	0.054291	0.657886	0.573643	0.000371	-0.202189	0.776327	0.581370	0.150177	-0.163481	0.511187	0.524744	-0.655220
16	0.150998	0.079084	-0.023281	0.494307	0.479602	-0.082345	-0.265499	0.573196	0.426339	-0.008061	-0.291269	0.310419	0.293180	-0.308922

Creating correlation matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
Apps	1.000000	0.955307	0.896883	0.321342	0.364491	0.861002	0.519823	0.065337	0.187475	0.236138	0.229948	0.463924
Accept	0.955307	1.000000	0.935277	0.223298	0.273681	0.897034	0.572691	-0.005002	0.119586	0.208705	0.256346	0.427341
Enroll	0.896883	0.935277	1.000000	0.171756	0.230434	0.967302	0.641595	-0.155655	-0.023846	0.202057	0.339348	0.381540
Top10perc	0.321342	0.223298	0.171756	1.000000	0.913875	0.111215	-0.180009	0.562160	0.357366	0.153452	-0.116730	0.544048
Top25perc	0.364491	0.273681	0.230434	0.913875	1.000000	0.181196	-0.099295	0.489569	0.330987	0.169761	-0.086810	0.551461
F.Undergrad	0.861002	0.897034	0.967302	0.111215	0.181196	1.000000	0.696130	-0.226166	-0.054476	0.207879	0.359783	0.361564
P.Undergrad	0.519823	0.572691	0.641595	-0.180009	-0.099295	0.696130	1.000000	-0.354216	-0.067638	0.122529	0.344053	0.127663
Outstate	0.065337	-0.005002	-0.155655	0.562160	0.489569	-0.226166	-0.354216	1.000000	0.655489	0.005110	-0.325609	0.391321
Room.Board	0.187475	0.119586	-0.023846	0.357366	0.330987	-0.054476	-0.067638	0.655489	1.000000	0.108924	-0.219554	0.341469
Books	0.236138	0.208705	0.202057	0.153452	0.169761	0.207879	0.122529	0.005110	0.108924	1.000000	0.239863	0.136390
Personal	0.229948	0.256346	0.339348	-0.116730	-0.086810	0.359783	0.344053	-0.325609	-0.219554	0.239863	1.000000	-0.011684
PhD	0.463924	0.427341	0.381540	0.544048	0.551461	0.361564	0.127663	0.391321	0.341469	0.136390	-0.011684	1.000000
Terminal	0.434478	0.403409	0.354379	0.506748	0.527654	0.335054	0.122152	0.412579	0.379270	0.159318	-0.031971	0.862928
S.F.Ratio	0.126411	0.188506	0.274269	-0.387926	-0.297233	0.324504	0.370607	-0.573683	-0.376430	-0.008536	0.173913	-0.129390
perc.alumni	-0.101158	-0.165516	-0.222723	0.455797	0.416832	-0.285457	-0.419334	0.565736	0.272393	-0.042832	-0.305753	0.248877
Expend	0.242935	0.161808	0.054221	0.657039	0.572905	0.000371	-0.201929	0.775328	0.580622	0.149983	-0.163271	0.510529
Grad.Rate	0.150803	0.078982	-0.023251	0.493670	0.478985	-0.082239	-0.265158	0.572458	0.425790	-0.008051	-0.290894	0.310019

"Covariance" indicates the direction of the linear relationship between variables. "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance.

Comment:

From this analysis establishes the fact that "standardizing the data-set and then computing the covariance and correlation matrices will yield the same results".

Q 2.5

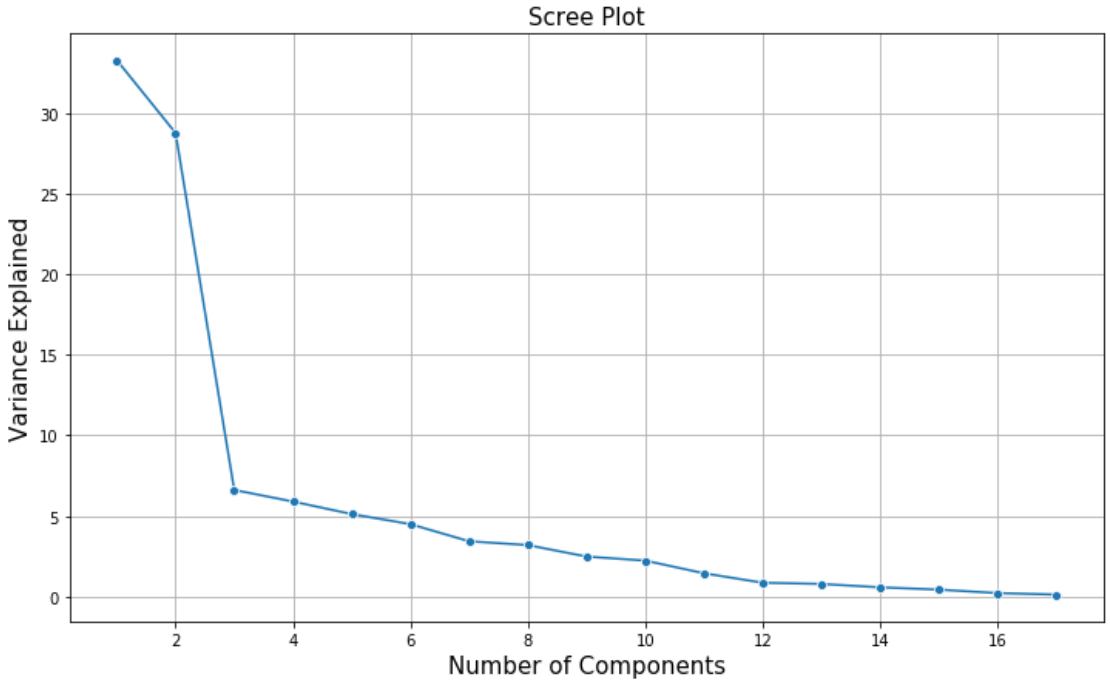
Step3 Creating covariance matrix, eigen values and eigen vectors:

```
Covariance Matrix
%s [[ 1.00128866e+00  9.56537704e-01  8.98039052e-01  3.21756324e-01
   3.64960691e-01  8.62111140e-01  5.20492952e-01  6.54209711e-02
   1.87717056e-01  2.36441941e-01  2.30243993e-01  4.64521757e-01
   4.35037784e-01  1.26573895e-01 -1.01288006e-01  2.43248206e-01
   1.50997775e-01]
 [ 9.56537704e-01  1.00128866e+00  9.36482483e-01  2.23586208e-01
   2.74033187e-01  8.98189799e-01  5.73428908e-01 -5.00874847e-03
   1.19740419e-01  2.08974091e-01  2.56676290e-01  4.27891234e-01
   4.03929238e-01  1.88748711e-01 -1.65728801e-01  1.62016688e-01
   7.90839722e-02]
 [ 8.98039052e-01  9.36482483e-01  1.00128866e+00  1.71977357e-01
   2.30730728e-01  9.68548601e-01  6.42421828e-01 -1.55856056e-01
  -2.38762560e-02  2.02317274e-01  3.39785395e-01  3.82031198e-01
   3.54835877e-01  2.74622251e-01 -2.23009677e-01  5.42906862e-02
  -2.32810071e-02]
 [ 3.21756324e-01  2.23586208e-01  1.71977357e-01  1.00128866e+00
   9.15052977e-01  1.11358019e-01 -1.80240778e-01  5.62884044e-01
   3.57826139e-01  1.53650150e-01 -1.16880152e-01  5.44748764e-01
   5.07401238e-01 -3.88425719e-01  4.56384036e-01  6.57885921e-01
   4.94306540e-01]
 [ 3.64960691e-01  2.74033187e-01  2.30730728e-01  9.15052977e-01
   1.00128866e+00  1.81429267e-01 -9.94231153e-02  4.90200034e-01
   3.31413314e-01  1.69979808e-01 -8.69219644e-02  5.52172085e-01
   5.28333659e-01 -2.97616423e-01  4.17369123e-01  5.73643193e-01
   4.79601950e-01]
 -
```

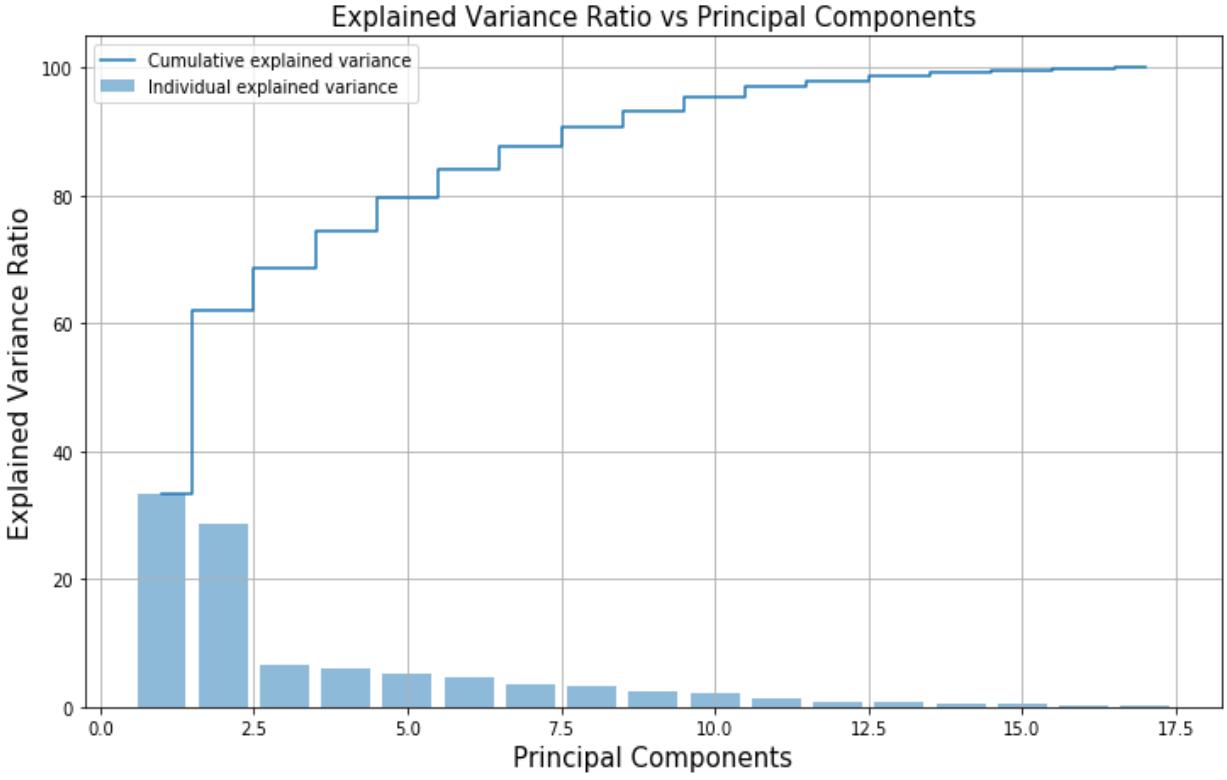
```
Eigen Values
%s [5.6625219  4.89470815 1.12636744 1.00397659 0.87218426 0.7657541
0.58491404 0.5445048  0.42352336 0.38101777 0.24701456 0.02239369
0.03789395 0.14726392 0.13434483 0.09883384 0.07469003]
```

```
Eigen Vectors
%s [[-2.62171542e-01  3.14136258e-01  8.10177245e-02 -9.87761685e-02
  -2.19898081e-01  2.18800617e-03 -2.83715076e-02 -8.99498102e-02
   1.30566998e-01 -1.56464458e-01 -8.62132843e-02  1.82169814e-01
  -5.99137640e-01  8.99775288e-02  8.88697944e-02  5.49428396e-01
   5.41453698e-03]
 [-2.30562461e-01  3.44623583e-01  1.07658626e-01 -1.18140437e-01
  -1.89634940e-01 -1.65212882e-02 -1.29584896e-02 -1.37606312e-01
   1.42275847e-01 -1.49209799e-01 -4.25899061e-02 -3.91041719e-01
   6.61496927e-01  1.58861886e-01  4.37945938e-02  2.91572312e-01
   1.44582845e-02]
 [-1.89276397e-01  3.82813322e-01  8.55296892e-02 -9.30717094e-03
  -1.62314818e-01 -6.80794143e-02 -1.52403625e-02 -1.44216938e-01
   5.08712481e-02 -6.48997860e-02 -4.38408622e-02  7.16684935e-01
   2.33235272e-01 -3.53988202e-02 -6.19241658e-02 -4.17001280e-01
  -4.97908902e-02]
 [-3.38874521e-01 -9.93191661e-02 -7.88293849e-02  3.69115031e-01
  -1.57211016e-01 -8.88656824e-02 -2.57455284e-01  2.89538833e-01
  -1.22467790e-01 -3.58776186e-02  1.77837341e-03 -5.62053913e-02
   2.21448729e-02 -3.92277722e-02  6.99599977e-02  8.79767299e-03
  -7.23645373e-01]
 [-3.34690532e-01 -5.95055011e-02 -5.07938247e-02  4.16824361e-01
  -1.44449474e-01 -2.76268979e-02 -2.39038849e-01  3.45643551e-01
  -1.93936316e-01  6.41786425e-03 -1.02127328e-01  1.96735274e-02
   3.22646978e-02  1.45621999e-01 -9.70282598e-02 -1.07779150e-02
   6.55464648e-01]
```

Q 2.6

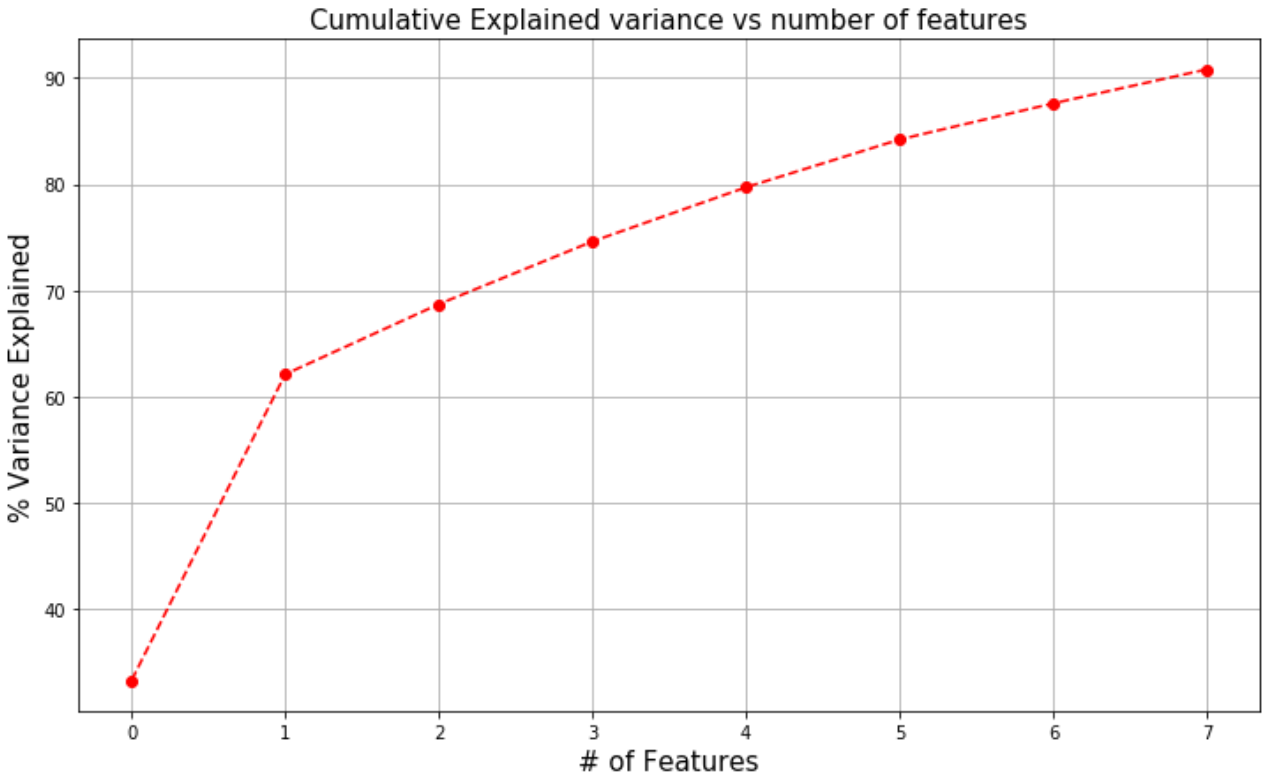


- Visually we can observe that there is steep drop in variance explained with increase in number of PC's.
- We will proceed with 8 components here. But its depending on requirement, here 90% variance is taken.



Q 2.7

```
Cumulative Variance Explained [ 33.26608367  62.02142867  68.63859223  74.53673619  79.66062886
  84.15926753  87.59551019  90.79435736  93.28246491  95.52086136
  96.97201814  97.83716159  98.62640821  99.20703552  99.64582321
  99.86844192 100.        ]
```



- The Cumulative % gives the percentage of variance accounted for by the n components. For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components.
- It helps in deciding the number of components by selecting the components which explained the high variance.
- In the above array we see that the first feature explains 32.5% of the variance within our data set while the first two explain 61.2 and so on.
- If we employ 8 features we capture ~ 90% of the variance within the dataset, thus we gain very little by implementing an additional feature.
- Thus we have created the Dataframe with Dimension Reduction Technique using Principal Component analysis and the 18 variable are reduced to 8 variable considering the 90% of variance in the dataset.

Names	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8
Abilene Christian University	-1.602499	0.993683	0.030045	-1.008422	-0.366886	-0.697476	0.710616	0.895167
Adelphi University	-1.804675	-0.070415	2.122128	3.138941	2.453212	0.994859	-0.396083	0.259664
Adrian College	-1.608283	-1.382792	-0.501513	-0.036373	0.765997	-1.026237	-0.165311	-0.408818
Agnes Scott College	2.803644	-3.367395	0.367768	-0.632914	-1.192601	-1.457080	-1.199862	0.357938
Alaska Pacific University	-2.200868	-0.099348	3.122523	0.657707	-1.828044	0.140915	-1.963228	-0.151893
...	...	...	...	...	...	...	...	...
Worcester State College	-3.395392	1.995628	-0.744776	0.800067	-0.342732	0.573074	-0.171927	0.098249
Xavier University	0.319750	-0.314944	0.013597	0.653856	0.462527	0.741736	0.778473	-0.271679
Xavier University of Louisiana	-0.576883	0.017798	0.322160	-0.587259	0.175225	0.504043	-1.458352	-0.289147
Yale University	6.570952	-1.184930	1.325966	0.077707	1.368517	-0.822746	1.201326	0.005740
York College of Pennsylvania	-0.477393	1.043947	-1.425438	-1.300274	0.720918	1.051810	1.073087	0.605417

Q 2.7

Business implication:

1. PCA removes Correlated Features: It is very common that you get more features in a dataset. After implementing the PCA on our dataset, all the Principal Components are independent of one another. There is no correlation among them, basically it removes the multicollinearity.

2. Improves Algorithm Performance: With so many dimensions, the performance of your algorithm will drastically degrade. PCA is a very common way to speed up your Machine Learning algorithm by getting rid of correlated variables which don't contribute in any decision making. The training time of the algorithms reduces significantly with less number of features. So, if the input dimensions are too high, then using PCA to speed up the algorithm is a reasonable choice.

3. Reduces Overfitting: Overfitting mainly occurs when there are too many variables in the dataset. So, PCA helps in overcoming the overfitting issue by reducing the number of features.

4. Improves Visualization: It is very hard to visualize and understand the data in high dimensions. PCA transforms a high dimensional data to low dimensional data (2 dimension) so that it can be visualized easily. We can use 2D Scree Plot to see which Principal Components result in high variance and have more impact as compared to other Principal Components.