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

EDA on the given dataset Fever.csv

1)Checking the top 5 records

	A	B	Volunteer	Relief
0	1	1	1	2.4
1	1	1	2	2.7
2	1	1	3	2.3
3	1	1	4	2.5
4	1	2	1	4.6

2) Checking the shape and information of the data frame

```
Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 A 36 non-null       int64

1 B 36 non-null       int64

2 Volunteer 36 non-null int64

3 Relief 36 non-null float64
```

3) Checking the summary of the data frame

	A	B	Volunteer	Relief
count	36.000000	36.000000	36.000000	36.000000
mean	2.000000	2.000000	2.500000	7.183333
std	0.828079	0.828079	1.133893	3.272090
min	1.000000	1.000000	1.000000	2.300000
25%	1.000000	1.000000	1.750000	4.675000
50%	2.000000	2.000000	2.500000	6.000000
75%	3.000000	3.000000	3.250000	9.325000
max	3.000000	3.000000	4.000000	13.500000

4) Checking distinct values of Design of A and B and Volunteer

For Variable A

```
3 12

2 12

1 12

Name: A, dtype: int64
```

For Variable B

```
3 12

2 12

1 12

Name: B, dtype: int64
```

For Variable Volunteer

```
4 9

3 9

2 9

1 9 Name: Volunteer, dtype: int64
```

1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually.

Answer

Null and Alternate hypothesis of Compound A:

Null Hypothesis *H*0 : u1 = u2 = u3 # The mean hours of relief provided by ingrident A is same at all three level.

Alternate Hypothesis H0: u1 = u2 != u3 # The mean hours of relief provided by ingrident A is different in at least one level

Null and Alternate hypothesis of Compound B:

Null Hypothesis *H*0 : u1 = u2 = u3 # The mean hours of relief provided by ingrident A is same at all three level.

Alternate Hypothesis H0: u1 = u2 != u3 # The mean hours of relief provided by ingrident A is different in at least one level

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.

	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

#since the p value is less than the significance level, we can reject the null hypothesis and states that there is a difference in the mean hours of relief provided by ingredient A,

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.

Answer

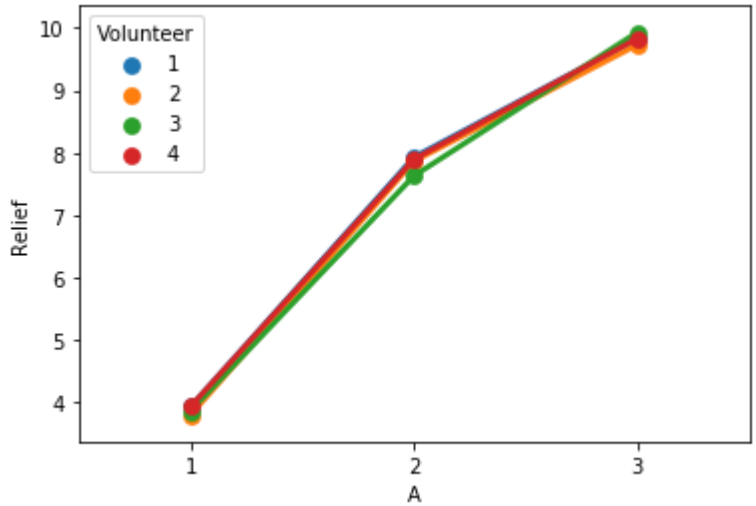
	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

#since the p value is less than the significance level, we can reject the null hypothesis and states that there is a difference in the mean hours of relief provided by ingredient B.

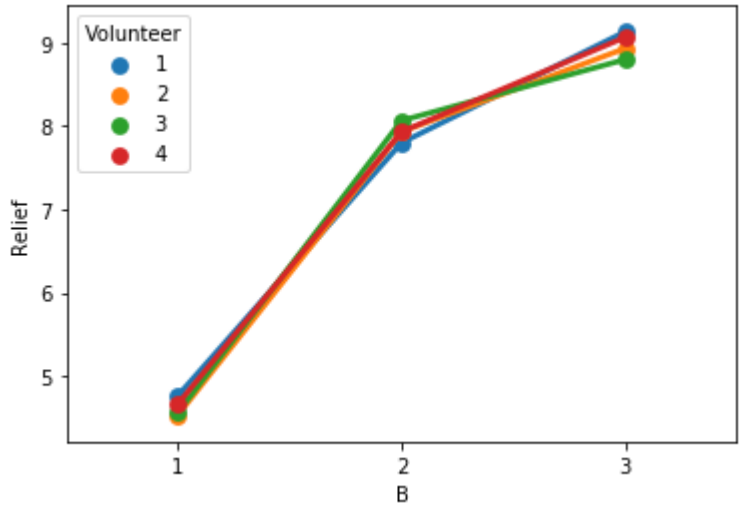
1.4) Analyse the effects of one variable on another with the help of an interaction plot.¶

What is an interaction between two treatments?

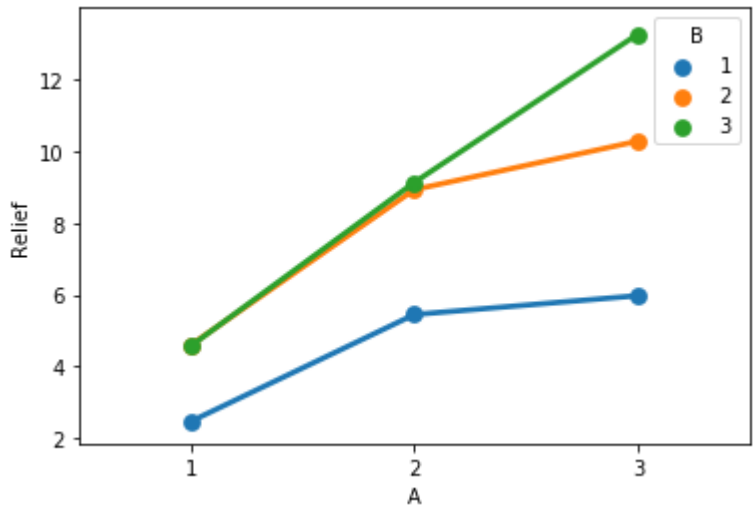
Answer



As we can see from the figure above by increasing the level of Compound A. Relief rate increases. Therefore we can say that Compound A is significant cause of Relief.



As we can see from the figure above by increasing the level of Compound B. Relief rate increases. Therefore we can say that Compound B is significant cause of Relief.



1) As we can see from the figure above
2) we can say that level 1 of Compound A and level 1 of Compound B, gives best result as we can get faster relief from hay fever

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

Answer

	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

We can see from the Anova table above we can state:

1. As p-value of Compound A is less than 0.05, we can say Compound A is a significant cause of Relief

2. As p-value of Compound B is less than 0.05, we can say Compound B is a significant cause of Relief

3. As A and B interaction is less than 0.05, there seems to be statistical interaction.

1.6) Mention the business implications of performing ANOVA for this particular case study

1. We can see if the level of Compound A increases, the Relief rate increases which is not desirable from the business perspective.

2. We can see if the level of Compound B increases, the Relief rate increases which is not desirable from the business perspective.

3. From interaction in two way Anova we can see that there is statistical interaction with compound A & B. Relief varies with interaction of compound A & B

4. As we can see from the interaction plot above Level 1 of Compound A & Level 1 of Compound B, gives a value between 2 and 4 hours of Relief time which is minimum and desirable.

Problem 2:¶

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

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

Answer

1)Checking the top 5 records

	0	1	2	3	4	5	6	7	8	9
Names	Abilene Christian University	Adelphi University	Adrian College	Agnes Scott College	Alaska Pacific University	Albertson College	Albertus Magnus College	Albion College	Albright College	Alderson-Broaddus College
Apps	1660	2186	1428	417	193	587	353	1899	1038	582
Accept	1232	1924	1097	349	146	479	340	1720	839	498
Enroll	721	512	336	137	55	158	103	489	227	172
Top10perc	23	16	22	60	16	38	17	37	30	21
Top25perc	52	29	50	89	44	62	45	68	63	44
F.Undergrad	2885	2683	1036	510	249	678	416	1594	973	799
P.Undergrad	537	1227	99	63	869	41	230	32	306	78
Outstate	7440	12280	11250	12960	7560	13500	13290	13868	15595	10468
Room.Board	3300	6450	3750	5450	4120	3335	5720	4826	4400	3380
Books	450	750	400	450	800	500	500	450	300	660
Personal	2200	1500	1165	875	1500	675	1500	850	500	1800
PhD	70	29	53	92	76	67	90	89	79	40
Terminal	78	30	66	97	72	73	93	100	84	41
S.F.Ratio	18.1	12.2	12.9	7.7	11.9	9.4	11.5	13.7	11.3	11.5
perc.alumni	12	16	30	37	2	11	26	37	23	15
Expend	7041	10527	8735	19016	10922	9727	8861	11487	11644	8991
Grad.Rate	60	56	54	59	15	55	63	73	80	52

2) Checking the shape and information of the data frame

```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Names        777 non-null    object
 1   Apps         777 non-null    int64
 2   Accept       777 non-null    int64
 3   Enroll       777 non-null    int64
 4   Top10perc    777 non-null    int64
 5   Top25perc    777 non-null    int64
 6   F.Undergrad  777 non-null    int64
 7   P.Undergrad  777 non-null    int64
 8   Outstate     777 non-null    int64
 9   Room.Board   777 non-null    int64
 10  Books        777 non-null    int64
 11  Personal     777 non-null    int64
 12  PhD          777 non-null    int64
 13  Terminal     777 non-null    int64
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64
 16  Expend       777 non-null    int64
 17  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16), object(1)
```

3) Summary of the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	University of Southern Mississippi	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777	NaN	NaN	NaN	3001.64	3870.2	81	776	1558	3624	48094
Accept	777	NaN	NaN	NaN	2018.8	2451.11	72	604	1110	2424	26330
Enroll	777	NaN	NaN	NaN	779.973	929.176	35	242	434	902	6392
Top10perc	777	NaN	NaN	NaN	27.5586	17.6404	1	15	23	35	96
Top25perc	777	NaN	NaN	NaN	55.7967	19.8048	9	41	54	69	100
F.Undergrad	777	NaN	NaN	NaN	3699.91	4850.42	139	992	1707	4005	31643
P.Undergrad	777	NaN	NaN	NaN	855.299	1522.43	1	95	353	967	21836
Outstate	777	NaN	NaN	NaN	10440.7	4023.02	2340	7320	9990	12925	21700
Room.Board	777	NaN	NaN	NaN	4357.53	1096.7	1780	3597	4200	5050	8124
Books	777	NaN	NaN	NaN	549.381	165.105	96	470	500	600	2340
Personal	777	NaN	NaN	NaN	1340.64	677.071	250	850	1200	1700	6800
PhD	777	NaN	NaN	NaN	72.6602	16.3282	8	62	75	85	103
Terminal	777	NaN	NaN	NaN	79.7027	14.7224	24	71	82	92	100
S.F.Ratio	777	NaN	NaN	NaN	14.0897	3.95835	2.5	11.5	13.6	16.5	39.8
perc.alumni	777	NaN	NaN	NaN	22.7439	12.3918	0	13	21	31	64
Expend	777	NaN	NaN	NaN	9660.17	5221.77	3186	6751	8377	10830	56233
Grad.Rate	777	NaN	NaN	NaN	65.4633	17.1777	10	53	65	78	118

4) Check for null values

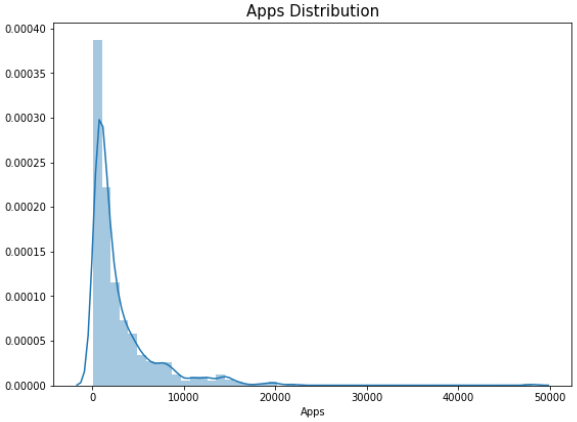
```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

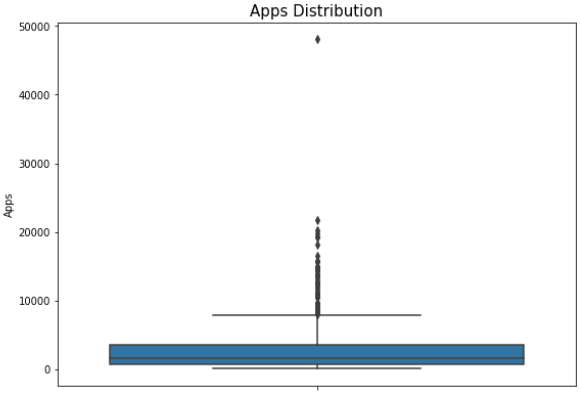
No null values

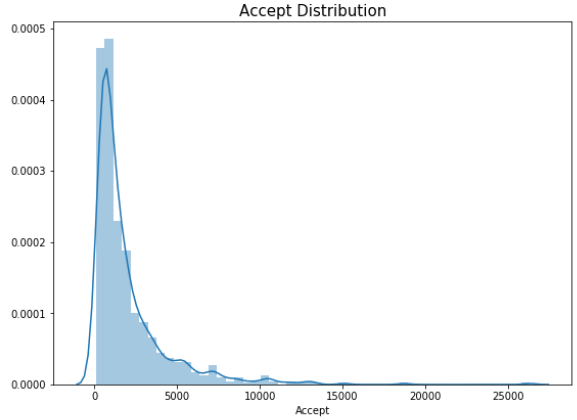
5) Check for duplicates

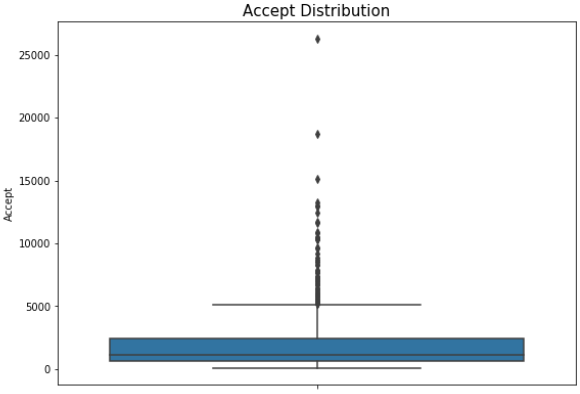
No duplicate values

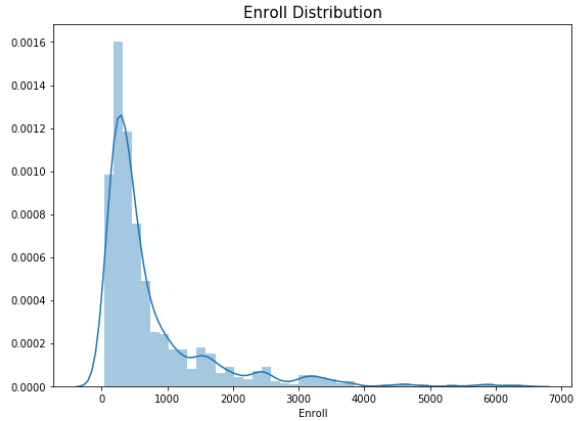
Uni-variate analysis(Of all the variables)

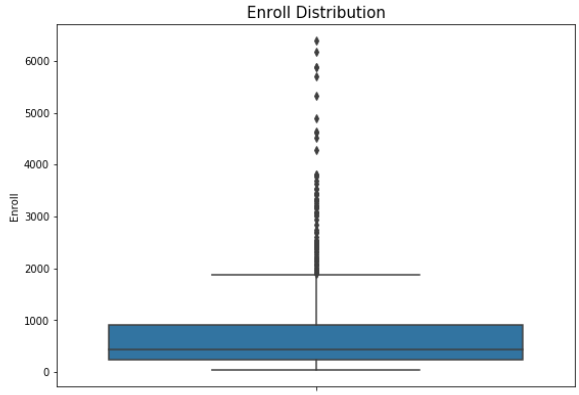


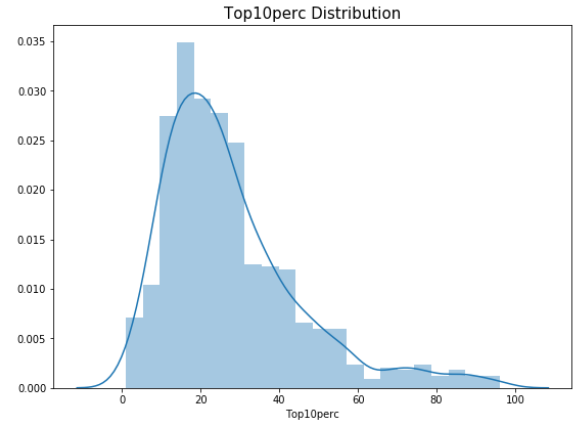


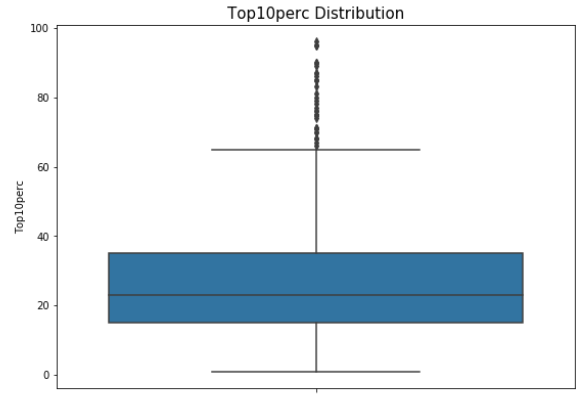


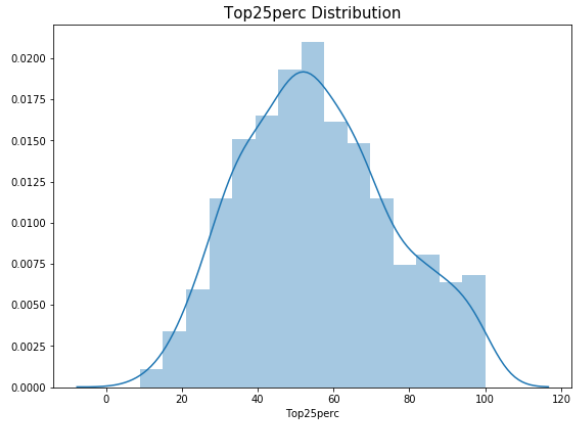


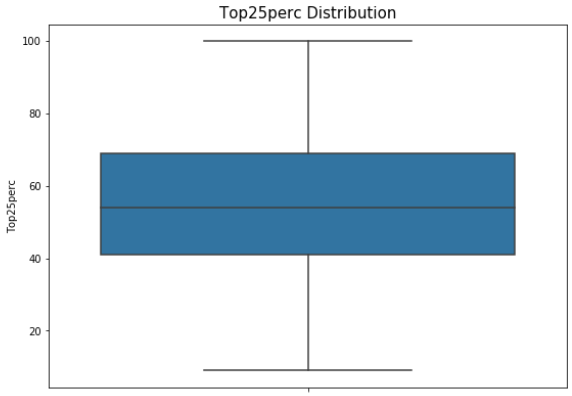


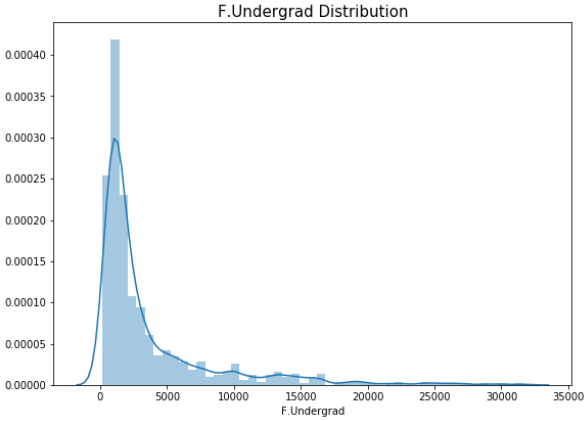


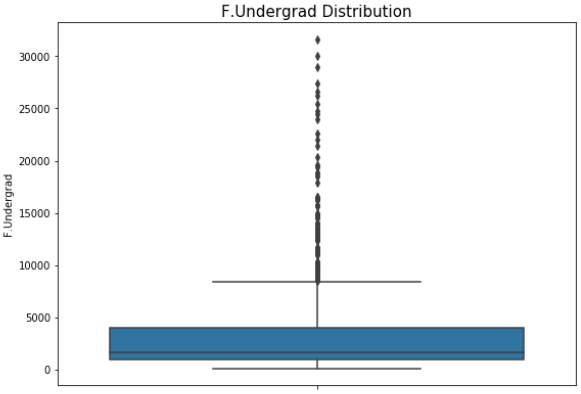


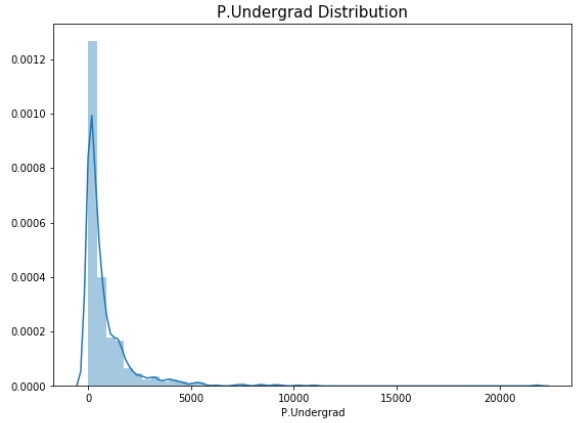


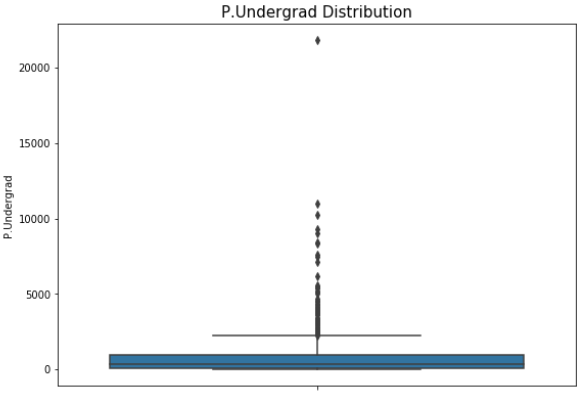

Top25perc Distribution
Top25perc

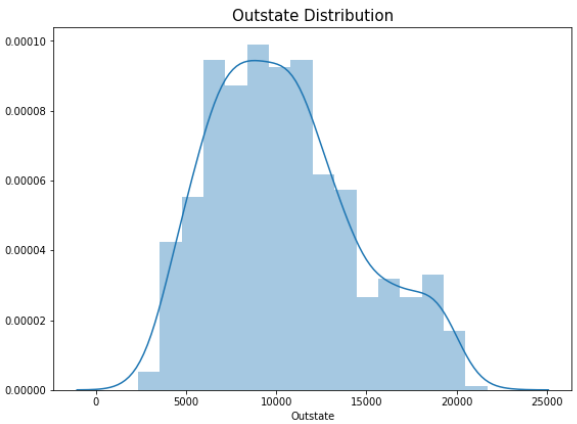

Top25perc Distribution
Top25perc

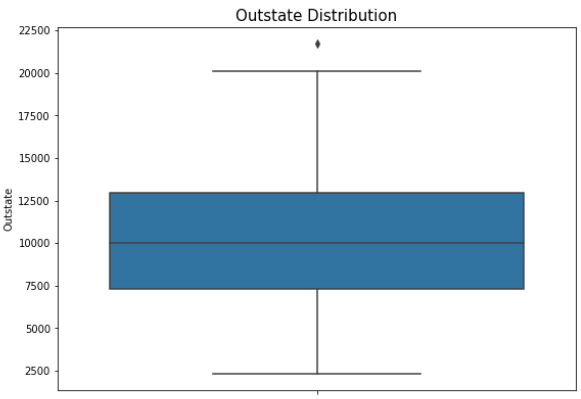

F.Undergrad Distribution
F.Undergrad

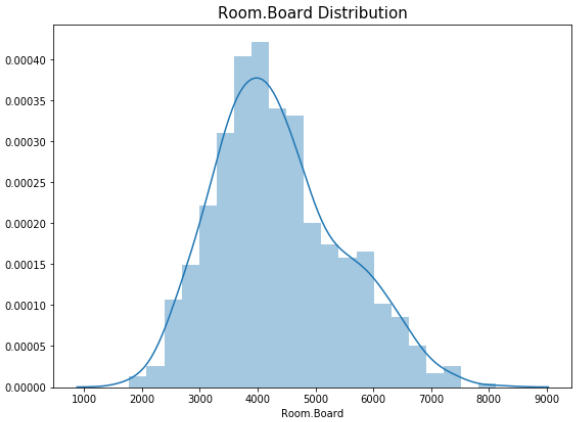

F.Undergrad Distribution
F.Undergrad

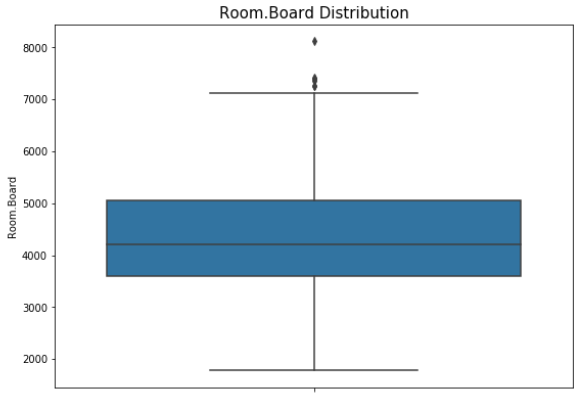

P.Undergrad Distribution
P.Undergrad

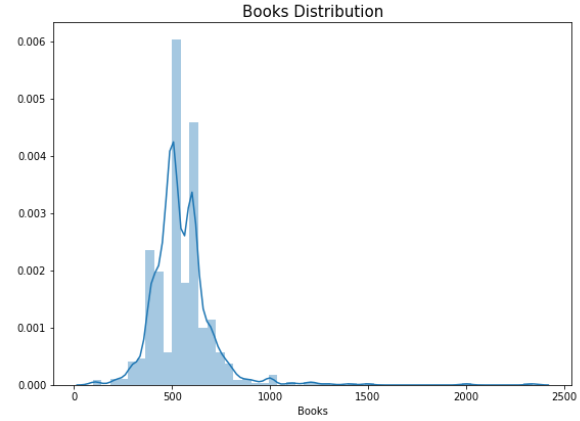

P.Undergrad Distribution
P.Undergrad

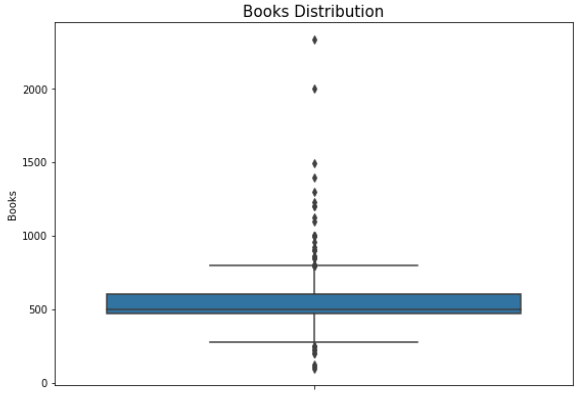

Outstate Distribution
Outstate

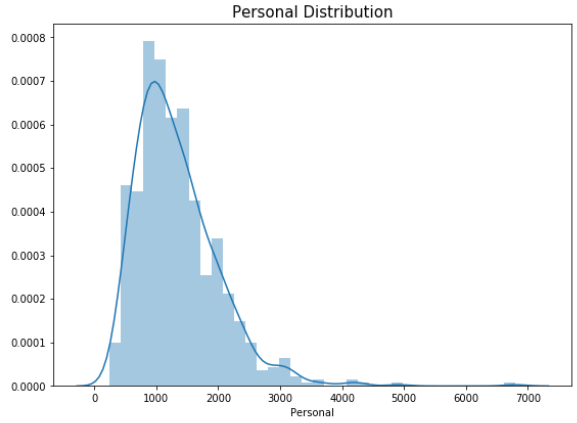

Outstate Distribution
Outstate

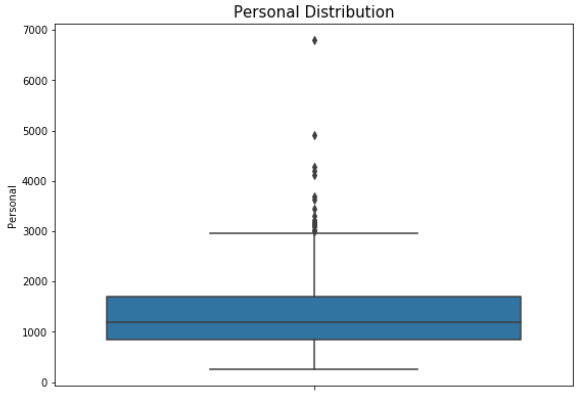

Room.Board Distribution
Room.Board

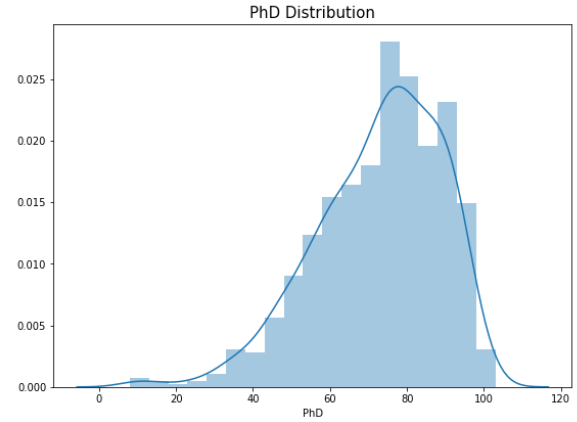

Room.Board Distribution
Room.Board

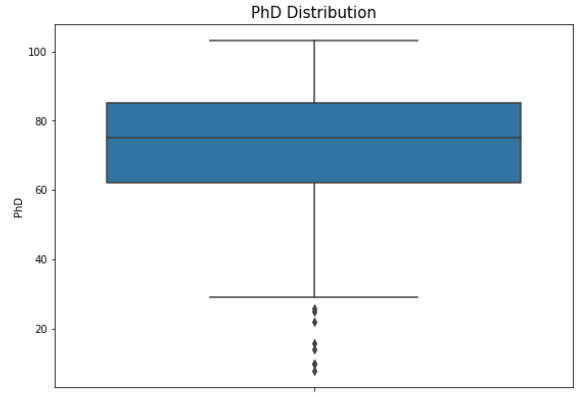

Books Distribution
Books

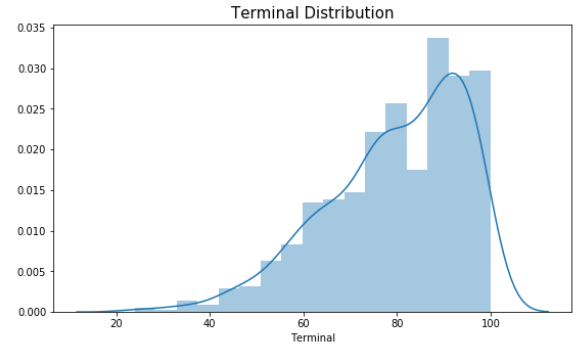

Books Distribution
Books

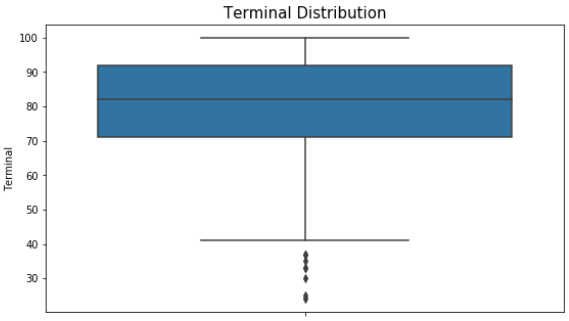

Personal Distribution
Personal

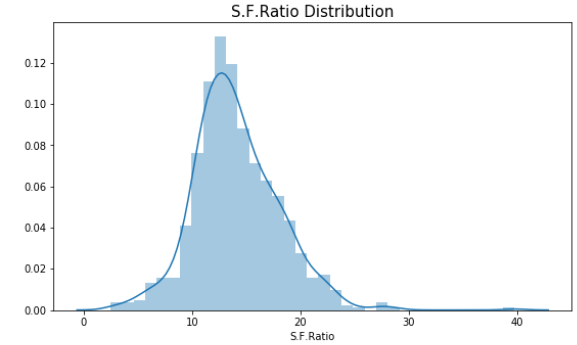

Personal Distribution
Personal

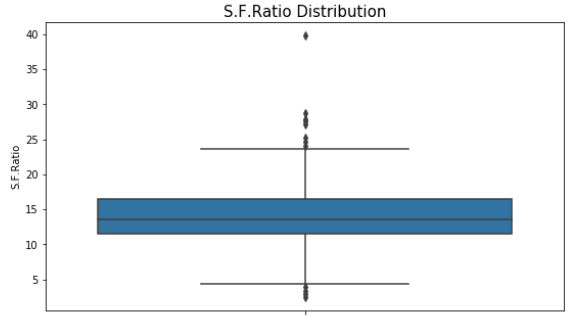

PhD Distribution
PhD

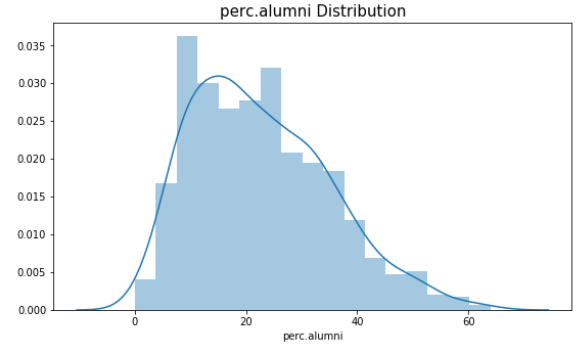

PhD Distribution
PhD

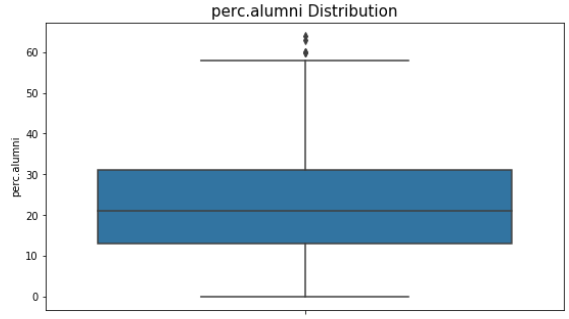

Terminal Distribution
Terminal

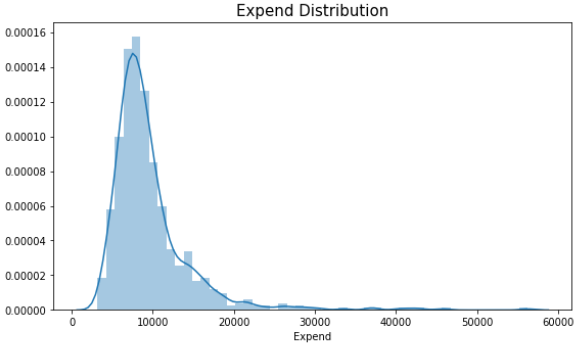

Terminal Distribution
Terminal

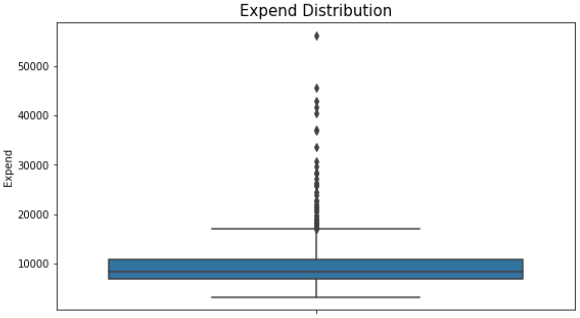

S.F.Ratio Distribution
S.F.Ratio

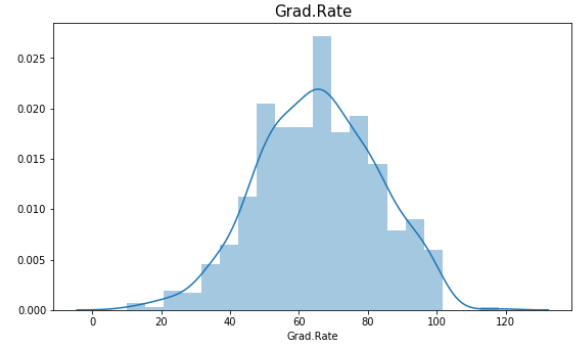

S.F.Ratio Distribution
S.F.Ratio

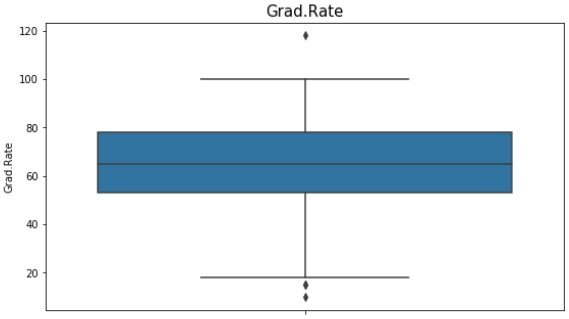

perc.alumni Distribution
perc.alumni


perc.alumni Distribution
perc.alumni


Expend Distribution
Expend


Expend Distribution
Expend


Grad.Rate
Grad.Rate


Grad.Rate
Grad.Rate

Inferences drawn from Univariate Analysis

1) We observe from the plots that the following variables have outliers Apps ,Accept,Enroll,Top10perc,Top25perc , F.Undergrad ,P.Undergrad,Books,Personal,PhD,Terminal,S.F.Ratio,perc,alumni,Expend,Grad.Rate.

2) Outstate and Room.Board ,Grad Rate variables have very few outliers.

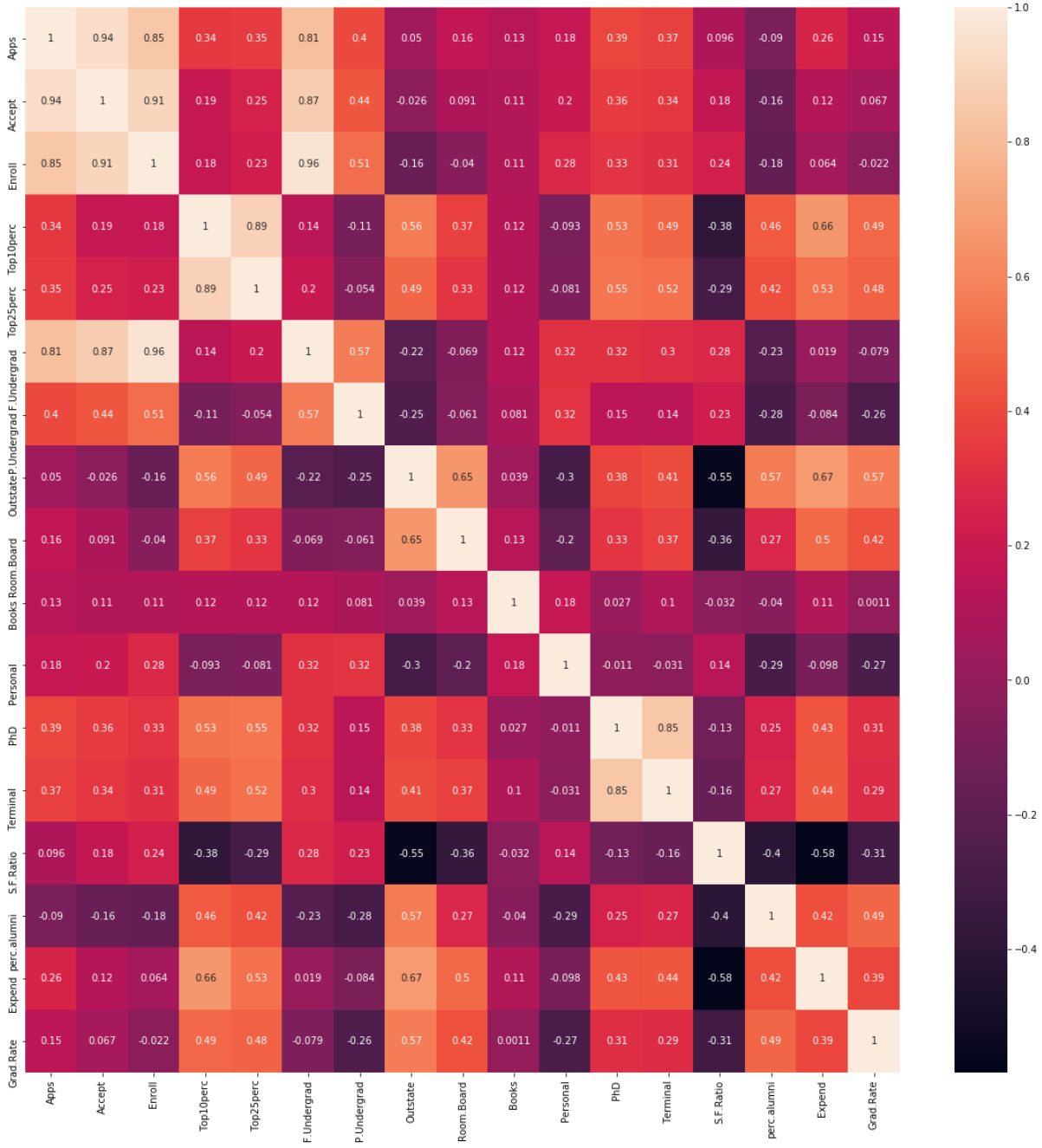
3) S.F.Ratio , Grad Rate varaibles shows some normal distribution with little skewness.

4) Almost all variables show skewness.

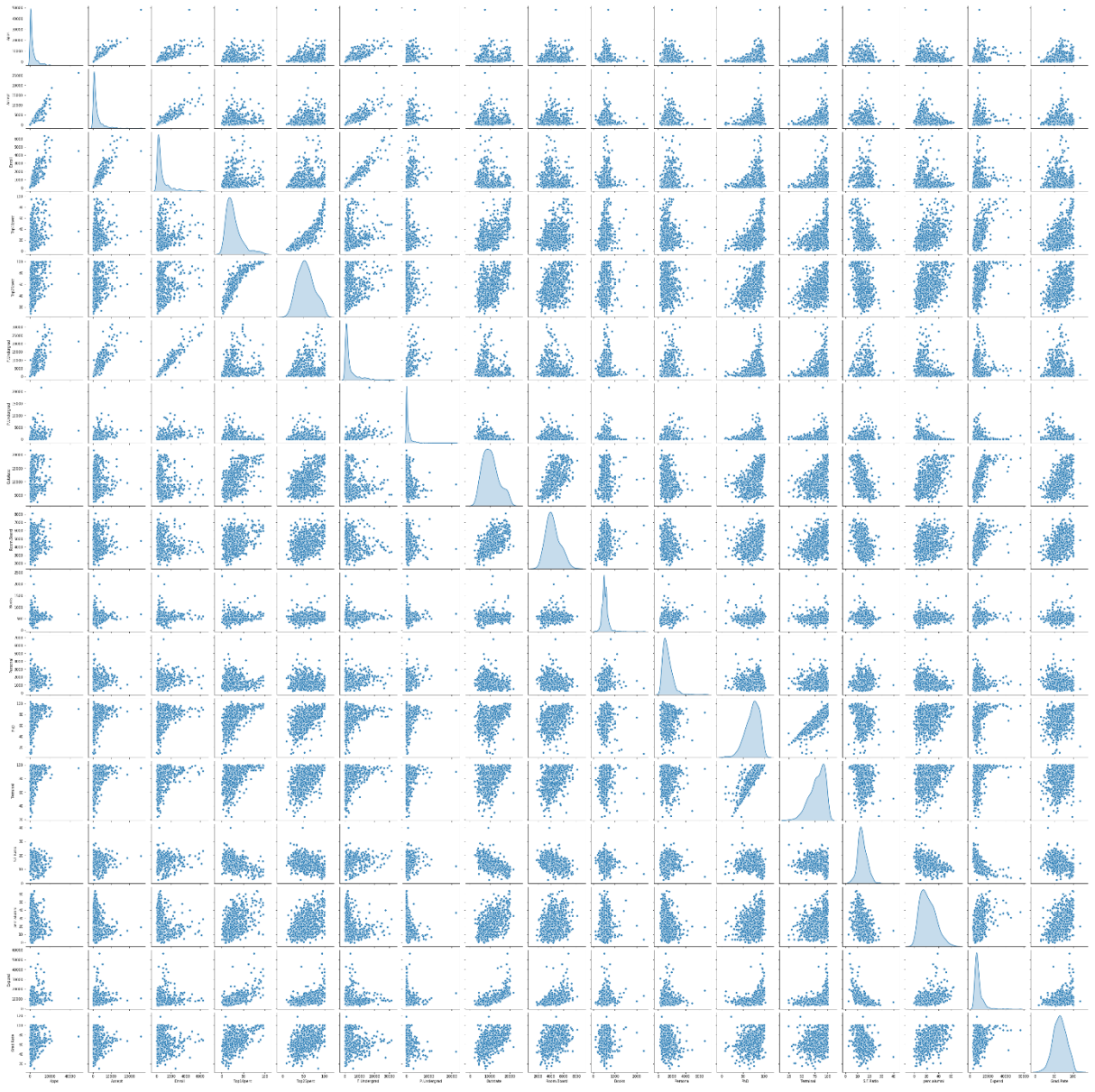
Multi-variate analysis(Of all the variables)

Answer

Plotting HeatMap for all the variables


Apps 1 0.94 0.85 0.34 0.35 0.81 0.4 0.05 0.16 0.13 0.18 0.39 0.37 0.096 -0.09 0.26 0.15
Accept 0.94 1 0.91 0.19 0.25 0.87 0.44 -0.026 0.091 0.11 0.2 0.36 0.34 0.18 -0.16 0.12 0.067
Enroll 0.85 0.91 1 0.18 0.23 0.96 0.51 -0.16 -0.04 0.11 0.28 0.33 0.31 0.24 -0.18 0.064 -0.022
Top10perc 0.34 0.19 0.18 1 0.89 0.14 -0.11 0.56 0.37 0.12 -0.093 0.53 0.49 -0.38 0.46 0.66 0.49
Top25perc 0.35 0.25 0.23 0.89 1 0.2 -0.054 0.49 0.33 0.12 -0.081 0.55 0.52 -0.29 0.42 0.53 0.48
F.Undergrad 0.81 0.87 0.96 0.14 0.2 1 0.57 -0.22 -0.069 0.12 0.32 0.32 0.3 0.28 -0.23 0.019 -0.079
P.Undergrad 0.4 0.44 0.51 -0.11 -0.054 0.57 1 -0.25 -0.061 0.081 0.32 0.15 0.14 0.23 -0.28 -0.084 -0.26
Outstate 0.05 -0.026 -0.16 0.56 0.49 -0.22 -0.25 1 0.65 0.039 -0.3 0.38 0.41 -0.55 0.57 0.67 0.57
Room.Board 0.16 0.091 -0.04 0.37 0.33 -0.069 -0.061 0.65 1 0.13 -0.2 0.33 0.37 -0.36 0.27 0.5 0.42
Books 0.13 0.11 0.11 0.12 0.12 0.12 0.081 0.039 0.13 1 0.18 0.027 0.1 -0.032 -0.04 0.11 0.0011
Personal 0.18 0.2 0.28 -0.093 -0.081 0.32 0.32 -0.3 -0.2 0.18 1 -0.011 -0.031 0.14 -0.29 -0.098 -0.27
PhD 0.39 0.36 0.33 0.53 0.55 0.32 0.15 0.38 0.33 0.027 -0.011 1 0.85 -0.13 0.25 0.43 0.31
Terminal 0.37 0.34 0.31 0.49 0.52 0.3 0.14 0.41 0.37 0.1 -0.031 0.85 1 -0.16 0.27 0.44 0.29
S.F.Ratio 0.096 0.18 0.24 -0.38 -0.29 0.28 0.23 -0.55 -0.36 -0.032 0.14 -0.13 -0.16 1 -0.4 -0.58 -0.31
perc.alumni -0.09 -0.16 -0.18 0.46 0.42 -0.23 -0.28 0.57 0.27 -0.04 -0.29 0.25 0.27 -0.4 1 0.42 0.49
Expend 0.26 0.12 0.064 0.66 0.53 0.019 -0.084 0.67 0.5 0.11 -0.098 0.43 0.44 -0.58 0.42 1 0.39
Grad.Rate 0.15 0.067 -0.022 0.49 0.48 -0.079 -0.26 0.57 0.42 0.0011 -0.27 0.31 0.29 -0.31 0.49 0.39 1
Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Plotting Pairplot for all the variables



Inferences drawn from Multivariate Analysis

We can see that many columns are co-related to each other or in other words the correlation exists for many columns and the highest is between Expend and outstate

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

Answer

Since Names column has unique values , so will remove it from the dataset.More over there is no point in adding ID for PCA.

	0	1	2	3	4
Apps	1660.0	2186.0	1428.0	417.0	193.0
Accept	1232.0	1924.0	1097.0	349.0	146.0
Enroll	721.0	512.0	336.0	137.0	55.0
Top10perc	23.0	16.0	22.0	60.0	16.0
Top25perc	52.0	29.0	50.0	89.0	44.0
F.Undergrad	2885.0	2683.0	1036.0	510.0	249.0
P.Undergrad	537.0	1227.0	99.0	63.0	869.0
Outstate	7440.0	12280.0	11250.0	12960.0	7560.0
Room.Board	3300.0	6450.0	3750.0	5450.0	4120.0
Books	450.0	750.0	400.0	450.0	795.0
Personal	2200.0	1500.0	1165.0	875.0	1500.0
PhD	70.0	29.0	53.0	92.0	76.0
Terminal	78.0	39.5	66.0	97.0	72.0
S.F.Ratio	18.1	12.2	12.9	7.7	11.9
perc.alumni	12.0	16.0	30.0	37.0	2.0
Expend	7041.0	10527.0	8735.0	16948.5	10922.0
Grad.Rate	60.0	56.0	54.0	59.0	15.5

	0	1	2	3	4	5	6	7	8	9
Apps	-0.376493	-0.159195	-0.472336	-0.889994	-0.982532	-0.819765	-0.916434	-0.277759	-0.633450	-0.821831
Accept	-0.337830	0.116744	-0.426511	-0.917871	-1.051221	-0.832474	-0.923783	-0.017264	-0.595991	-0.819993
Enroll	0.106380	-0.260441	-0.569343	-0.918613	-1.062533	-0.881755	-0.978287	-0.300809	-0.760651	-0.857183
Top10perc	-0.246780	-0.696290	-0.310996	2.129202	-0.696290	0.716456	-0.632075	0.652240	0.202730	-0.375212
Top25perc	-0.191827	-1.353911	-0.292878	1.677612	-0.596031	0.313426	-0.545505	0.616579	0.363952	-0.596031
F.Undergrad	-0.018769	-0.093626	-0.703966	-0.898889	-0.995610	-0.836632	-0.933723	-0.497184	-0.727312	-0.791792
P.Undergrad	-0.166083	0.797856	-0.777974	-0.828267	0.297726	-0.859001	-0.594966	-0.871574	-0.488793	-0.807312
Outstate	-0.746480	0.457762	0.201488	0.626954	-0.716623	0.761311	0.709061	0.852873	1.282569	0.006918
Room.Board	-0.968324	1.921680	-0.555466	1.004218	-0.216006	-0.936213	1.251933	0.431722	0.040884	-0.894927
Books	-0.776567	1.828605	-1.210762	-0.776567	2.219381	-0.342372	-0.342372	-0.776567	-2.079153	1.047053
Personal	1.438500	0.289289	-0.260691	-0.736792	0.289289	-1.065138	0.289289	-0.777836	-1.352441	0.781808
PhD	-0.174045	-2.745731	-1.240354	1.205884	0.202299	-0.362217	1.080436	1.017712	0.390471	-2.055766
Terminal	-0.123239	-2.785068	-0.952900	1.190391	-0.538069	-0.468931	0.913837	1.397806	0.291591	-2.681360
S.F.Ratio	1.070602	-0.489511	-0.304413	-1.679429	-0.568839	-1.229904	-0.674610	-0.092872	-0.727495	-0.674610
perc.alumni	-0.870466	-0.545726	0.590864	1.159159	-1.682316	-0.951651	0.266124	1.159159	0.022569	-0.626911
Expend	-0.630916	0.396097	-0.131845	2.287940	0.512468	0.160409	-0.094724	0.678923	0.725177	-0.056425
Grad.Rate	-0.319205	-0.552693	-0.669437	-0.377577	-2.916759	-0.611065	-0.144089	0.439631	0.848234	-0.786181

Inferences drawn from Scaling

Using normal scaling (Z score) we scale the variables in range of -1 to 1 in order to perform PCA. We can see now the data for all the variables now lies on a scale from -1 to 1 and the variation in data is not there anymore.

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

Answer

1. Covariance indicates the direction of the linear relationship between variables.

2. Correlation on the other hand measures both the strength and direction of the linear relationship between two Variables and it is quantified by the Correlation Coefficients

3. Correlation is a function of the covariance

4. We can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.

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

Answer

Before Scaling on removing the outliers -

```
Accept         0
Apps           0
Books          0
Enroll         0
Expend         0
F.Undergrad    0
Grad.Rate      0
Names          0
Outstate       0
P.Undergrad    0
Personal       0
PhD            0
Room.Board     0
S.F.Ratio      0
Terminal       0
Top10perc      0
Top25perc      0
perc.alumni    0
dtype: int64
```

After Scaling

```
Apps           0
Accept         0
Enroll         0
Top10perc     40
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      8
perc.alumni    6
Expend         0
Grad.Rate      0
dtype: int64
```

Inferences drawn on checking outliers after and before Scaling

We can say that on removing the outliers before scaling there were no outliers in any of the variables. But after scaling we see some outliers in three variables i.e. Top10perc , S.F.Ratio , perc.alumni.

2.5 Build the covariance matrix, eigenvalues and eigenvector.

Answer

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] [8.62111140e-01
8.98189799e-01 9.68548601e-01 1.11358019e-01 1.81429267e-01
1.00128866e+00 6.97027420e-01 -2.26457040e-01 -5.45459528e-02
2.08147257e-01 3.60246460e-01 3.62030390e-01 3.35485771e-01
3.24921933e-01 -2.85825062e-01 3.71119607e-04 -8.23447851e-02] [
5.20492952e-01 5.73428908e-01 6.42421828e-01 -1.80240778e-01 -
9.94231153e-02 6.97027420e-01 1.00128866e+00 -3.54672874e-01 -
6.77252009e-02 1.22686416e-01 3.44495974e-01 1.27827147e-01
1.22309141e-01 3.71084841e-01 -4.19874031e-01 -2.02189396e-01 -
2.65499420e-01] [6.54209711e-02 -5.00874847e-03 -1.55856056e-01
5.62884044e-01 4.90200034e-01 -2.26457040e-01 -3.54672874e-01
1.00128866e+00 6.56333564e-01 5.11656377e-03 -3.26028927e-01
3.91824814e-01 4.13110264e-01 -5.74421963e-01 5.66465309e-01
7.76326650e-01 5.73195743e-01] [1.87717056e-01 1.19740419e-01 -
2.38762560e-02 3.57826139e-01 3.31413314e-01 -5.45459528e-02 -

6.77252009e-02 6.56333564e-01 1.00128866e+00 1.09064551e-01 -
2.19837042e-01 3.41908577e-01 3.79759015e-01 -3.76915472e-01
2.72743761e-01 5.81370284e-01 4.26338910e-01] [2.36441941e-01
2.08974091e-01 2.02317274e-01 1.53650150e-01 1.69979808e-01
2.08147257e-01 1.22686416e-01 5.11656377e-03 1.09064551e-01
1.00128866e+00 2.40172145e-01 1.36566243e-01 1.59523091e-01 -
8.54689129e-03 -4.28870629e-02 1.50176551e-01 -8.06107505e-03] [
2.30243993e-01 2.56676290e-01 3.39785395e-01 -1.16880152e-01 -
8.69219644e-02 3.60246460e-01 3.44495974e-01 -3.26028927e-01 -
2.19837042e-01 2.40172145e-01 1.00128866e+00 -1.16986124e-02 -
3.20117803e-02 1.74136664e-01 -3.06146886e-01 -1.63481407e-01 -
2.91268705e-01] [4.64521757e-01 4.27891234e-01 3.82031198e-01
5.44748764e-01 5.52172085e-01 3.62030390e-01 1.27827147e-01
3.91824814e-01 3.41908577e-01 1.36566243e-01 -1.16986124e-02
1.00128866e+00 8.64040263e-01 -1.29556494e-01 2.49197779e-01
5.11186852e-01 3.10418895e-01] [4.35037784e-01 4.03929238e-01
3.54835877e-01 5.07401238e-01 5.28333659e-01 3.35485771e-01
1.22309141e-01 4.13110264e-01 3.79759015e-01 1.59523091e-01 -
3.20117803e-02 8.64040263e-01 1.00128866e+00 -1.51187934e-01
2.66375402e-01 5.24743500e-01 2.93180212e-01] [1.26573895e-01
1.88748711e-01 2.74622251e-01 -3.88425719e-01 -2.97616423e-01
3.24921933e-01 3.71084841e-01 -5.74421963e-01 -3.76915472e-01 -
8.54689129e-03 1.74136664e-01 -1.29556494e-01 -1.51187934e-01
1.00128866e+00 -4.12632056e-01 -6.55219504e-01 -3.08922187e-01] [-
1.01288006e-01 -1.65728801e-01 -2.23009677e-01 4.56384036e-01
4.17369123e-01 -2.85825062e-01 -4.19874031e-01 5.66465309e-01
2.72743761e-01 -4.28870629e-02 -3.06146886e-01 2.49197779e-01
2.66375402e-01 -4.12632056e-01 1.00128866e+00 4.63518674e-01
4.92040760e-01] [2.43248206e-01 1.62016688e-01 5.42906862e-02
6.57885921e-01 5.73643193e-01 3.71119607e-04 -2.02189396e-01
7.76326650e-01 5.81370284e-01 1.50176551e-01 -1.63481407e-01
5.11186852e-01 5.24743500e-01 -6.55219504e-01 4.63518674e-01
1.00128866e+00 4.15826026e-01] [1.50997775e-01 7.90839722e-02 -
2.32810071e-02 4.94306540e-01 4.79601950e-01 -8.23447851e-02 -
2.65499420e-01 5.73195743e-01 4.26338910e-01 -8.06107505e-03 -
2.91268705e-01 3.10418895e-01 2.93180212e-01 -3.08922187e-01
4.92040760e-01 4.15826026e-01 1.00128866e+00]]

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

Eigen Vectors %s [[-2.62171542e-01 3.14136258e-01 -8.10177245e-
02 9.87761685e-02 2.19898081e-01 -2.18800617e-03 2.83715076e-02
8.99498102e-02 -1.30566998e-01 1.56464458e-01 8.62132843e-02 -
1.82169814e-01 5.99137640e-01 -8.99775288e-02 -8.88697944e-02
5.49428396e-01 5.41453698e-03] [-2.30562461e-01 3.44623583e-01 -

1.07658626e-01 1.18140437e-01 1.89634940e-01 1.65212882e-02
1.29584896e-02 1.37606312e-01 -1.42275847e-01 1.49209799e-01
4.25899061e-02 3.91041719e-01 -6.61496927e-01 -1.58861886e-01 -
4.37945938e-02 2.91572312e-01 1.44582845e-02] [-1.89276397e-01
3.82813322e-01 -8.55296892e-02 9.30717094e-03 1.62314818e-01
6.80794143e-02 1.52403625e-02 1.44216938e-01 -5.08712481e-02
6.48997860e-02 4.38408622e-02 -7.16684935e-01 -2.33235272e-01
3.53988202e-02 6.19241658e-02 -4.17001280e-01 -4.97908902e-02] [-
3.38874521e-01 -9.93191661e-02 7.88293849e-02 -3.69115031e-01
1.57211016e-01 8.88656824e-02 2.57455284e-01 -2.89538833e-01
1.22467790e-01 3.58776186e-02 -1.77837341e-03 5.62053913e-02 -
2.21448729e-02 3.92277722e-02 -6.99599977e-02 8.79767299e-03 -
7.23645373e-01] [-3.34690532e-01 -5.95055011e-02 5.07938247e-02 -
4.16824361e-01 1.44449474e-01 2.76268979e-02 2.39038849e-01 -
3.45643551e-01 1.93936316e-01 -6.41786425e-03 1.02127328e-01 -
1.96735274e-02 -3.22646978e-02 -1.45621999e-01 9.70282598e-02 -
1.07779150e-02 6.55464648e-01] [-1.63293010e-01 3.98636372e-01 -
7.37077827e-02 1.39504424e-02 1.02728468e-01 5.16468727e-02
3.11751439e-02 1.08748900e-01 -1.45452749e-03 1.63981359e-04
3.49993487e-02 5.42774834e-01 3.67681187e-01 1.33555923e-01
8.71753137e-02 -5.70683843e-01 2.53059904e-02] [-2.24797091e-02
3.57550046e-01 -4.03568700e-02 2.25351078e-01 -9.56790178e-02
2.45375721e-02 1.00138971e-02 -1.23841696e-01 6.34774326e-01 -
5.46346279e-01 -2.52107094e-01 -2.95029745e-02 -2.62494456e-02 -
5.02487566e-02 -4.45537493e-02 1.46321060e-01 -3.97146972e-02] [-
2.83547285e-01 -2.51863617e-01 -1.49394795e-02 2.62975384e-01
3.72750885e-02 2.03860462e-02 -9.45370782e-02 -1.12721477e-02
8.36648339e-03 2.31799759e-01 -5.93433149e-01 -1.03393587e-03
8.14247697e-02 -5.60392799e-01 -6.72405494e-02 -2.11561014e-01 -
1.59275617e-03] [-2.44186588e-01 -1.31909124e-01 2.11379165e-02
5.80894132e-01 -6.91080879e-02 -2.37267409e-01 -9.45210745e-02 -
3.89639465e-01 2.20526518e-01 2.55107620e-01 4.75297296e-01 -
9.85725168e-03 -2.67779296e-02 1.07365653e-01 -1.77715010e-02 -
1.00935084e-01 -2.82578388e-02] [-9.67082754e-02 9.39739472e-02
6.97121128e-01 -3.61562884e-02 3.54056654e-02 -6.38604997e-01
1.11193334e-01 2.39817267e-01 -2.10246624e-02 -9.11624912e-02 -
4.35697999e-02 -4.36086500e-03 -1.04624246e-02 -5.16224550e-02 -
3.54343707e-02 -2.86384228e-02 -8.06259380e-03] [3.52299594e-02
2.32439594e-01 5.30972806e-01 -1.14982973e-01 -4.75358244e-04
3.81495854e-01 -6.39418106e-01 -2.77206569e-01 -1.73715184e-02
1.27647512e-01 -1.51627393e-02 1.08725257e-02 -4.54572099e-03 -
9.39409228e-03 1.18604404e-02 3.38197909e-02 1.42590097e-03] [-
3.26410696e-01 5.51390195e-02 -8.11134044e-02 -1.47260891e-01 -
5.50786546e-01 -3.34444832e-03 -8.92320786e-02 3.42628480e-02 -
1.66510079e-01 -1.00975002e-01 3.91865961e-02 -1.33146759e-02 -
1.25137966e-02 7.16590441e-02 -7.02656469e-01 -6.38096394e-02
8.31471932e-02] [-3.23115980e-01 4.30332048e-02 -5.89785929e-02 -
8.90079921e-02 -5.90407136e-01 -3.54121294e-02 -9.16985445e-02
9.03076644e-02 -1.12609034e-01 -8.60363025e-02 8.48575651e-02 -
7.38135022e-03 1.79275275e-02 -1.63820871e-01 6.62488717e-01
9.85019644e-02 -1.13374007e-01] [1.63151642e-01 2.59804556e-01 -
2.74150657e-01 -2.59486122e-01 -1.42842546e-01 -4.68752604e-01 -
1.52864837e-01 -2.42807562e-01 1.53685343e-01 4.70527925e-01 -
3.63042716e-01 -8.85797314e-03 -1.83059753e-02 2.39902591e-01

```
4.79006197e-02 6.19970446e-02 3.83160891e-03] [-1.86610828e-01 -
2.57092552e-01 -1.03715887e-01 -2.23982467e-01 1.28215768e-01 -
1.25669415e-02 -3.91400512e-01 5.66073056e-01 5.39235753e-01
1.47628917e-01 1.73918533e-01 2.40534190e-02 8.03169296e-05
4.89753356e-02 -3.58875507e-02 2.80805469e-02 -7.32598621e-03] [-
3.28955847e-01 -1.60008951e-01 1.84205687e-01 2.13756140e-01 -
2.24240837e-02 2.31562325e-01 1.50501305e-01 1.18823549e-01 -
2.42371616e-02 8.04154875e-02 -3.93722676e-01 -1.05658769e-02 -
5.60069250e-02 6.90417042e-01 1.26667522e-01 1.28739213e-01
1.45099786e-01] [-2.38822447e-01 -1.67523664e-01 -2.45335837e-01 -
3.61915064e-02 3.56843227e-01 -3.13556243e-01 -4.68641965e-01 -
1.80458508e-01 -3.15812873e-01 -4.88415259e-01 -8.72638706e-02
2.51028410e-03 -1.48410810e-02 1.59332164e-01 6.30737002e-02 -
7.09643331e-03 -3.29024228e-03]]
```

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

Answer

Explicit form of the first PC

```
Apps * (-2.62171542e-01) + Accept * (3.14136258e-01) + Enroll * (-
8.10177245e-02) + Top10perc * (9.87761685e-02) + Top25perc *
(2.19898081e-01) + F.Undergrad * (-2.18800617e-03) + P.Undergrad *
(2.83715076e-02) + Outstate * (8.99498102e-02) + Room.Board * (-
1.30566998e-01 ) + Books * (1.56464458e-01) + Personal * (8.62132843e-
02) + PhD * (-1.82169814e-01) + Terminal * (5.99137640e-01) + S.F.Ratio
* (-8.99775288e-02) + perc.alumni * (-8.88697944e-02) + Expend *
(5.49428396e-01) + Grad.Rate * (5.41453698e-03)
```

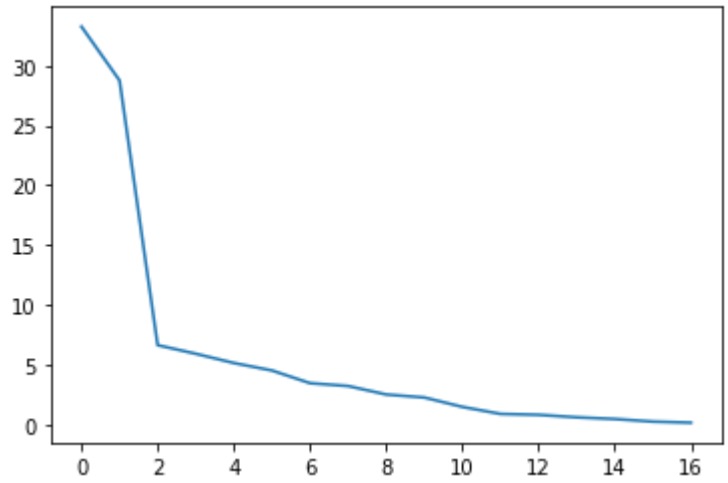
2.7) Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?¶

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

Answer

Cumulative Values [33.26608367 62.02142867 68.63859223 74.536
73619 79.66062886 84.15926753 87.59551019 90.79435736 93.28246491 95.52
086136 96.97201814 97.83716159 98.62640821 99.20703552 99.64582321 99.8
6844192 100.]

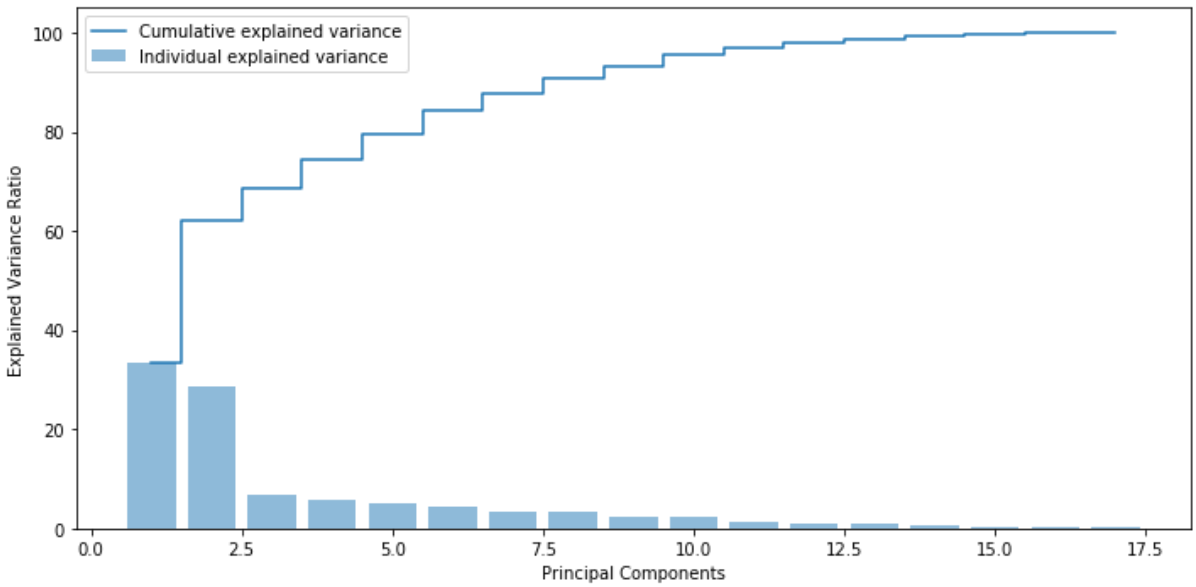
Plotting on Scree Plot



1) Visually we can observe that there is steep drop in variance explained with increase in number of PC's.

2) We will proceed with 5 components here. But depending on requirement 90% variation or 6 components will also do well.

Eigen Vectors indicate the direction of the new features space.



PCA

PCA components (taking 5 components) –

```
array([[ 2.62171542e-01, 2.30562461e-01, 1.89276397e-01, 3.38874521e-
01, 3.34690532e-01, 1.63293010e-01, 2.24797091e-02, 2.83547285e-01,
2.44186588e-01, 9.67082754e-02, -3.52299594e-02, 3.26410696e-01,
3.23115980e-01, -1.63151642e-01, 1.86610828e-01, 3.28955847e-01,
2.38822447e-01], [ 3.14136258e-01, 3.44623583e-01, 3.82813322e-01, -
9.93191661e-02, -5.95055011e-02, 3.98636372e-01, 3.57550046e-01, -
2.51863617e-01, -1.31909124e-01, 9.39739472e-02, 2.32439594e-01,
5.51390195e-02, 4.30332048e-02, 2.59804556e-01, -2.57092552e-01, -
1.60008951e-01, -1.67523664e-01], [-8.10177326e-02, -1.07658616e-01, -
8.55296879e-02, 7.88293855e-02, 5.07938249e-02, -7.37077862e-02, -
4.03568697e-02, -1.49394805e-02, 2.11379168e-02, 6.97121128e-01,
5.30972806e-01, -8.11134043e-02, -5.89785931e-02, -2.74150657e-01, -
1.03715887e-01, 1.84205687e-01, -2.45335836e-01], [ 9.87761891e-02,
1.18140414e-01, 9.30716539e-03, -3.69115032e-01, -4.16824362e-01,
1.39504530e-02, 2.25351077e-01, 2.62975387e-01, 5.80894131e-01, -
3.61562887e-02, -1.14982973e-01, -1.47260891e-01, -8.90079915e-02, -
2.59486123e-01, -2.23982467e-01, 2.13756138e-01, -3.61915069e-02], [
2.19898026e-01, 1.89634999e-01, 1.62314842e-01, 1.57211019e-01,
1.44449476e-01, 1.02728432e-01, -9.56790152e-02, 3.72750810e-02, -
6.91080854e-02, 3.54056664e-02, -4.75357868e-04, -5.50786545e-01, -
5.90407138e-01, -1.42842545e-01, 1.28215768e-01, -2.24240787e-02,
3.56843228e-01]])
```

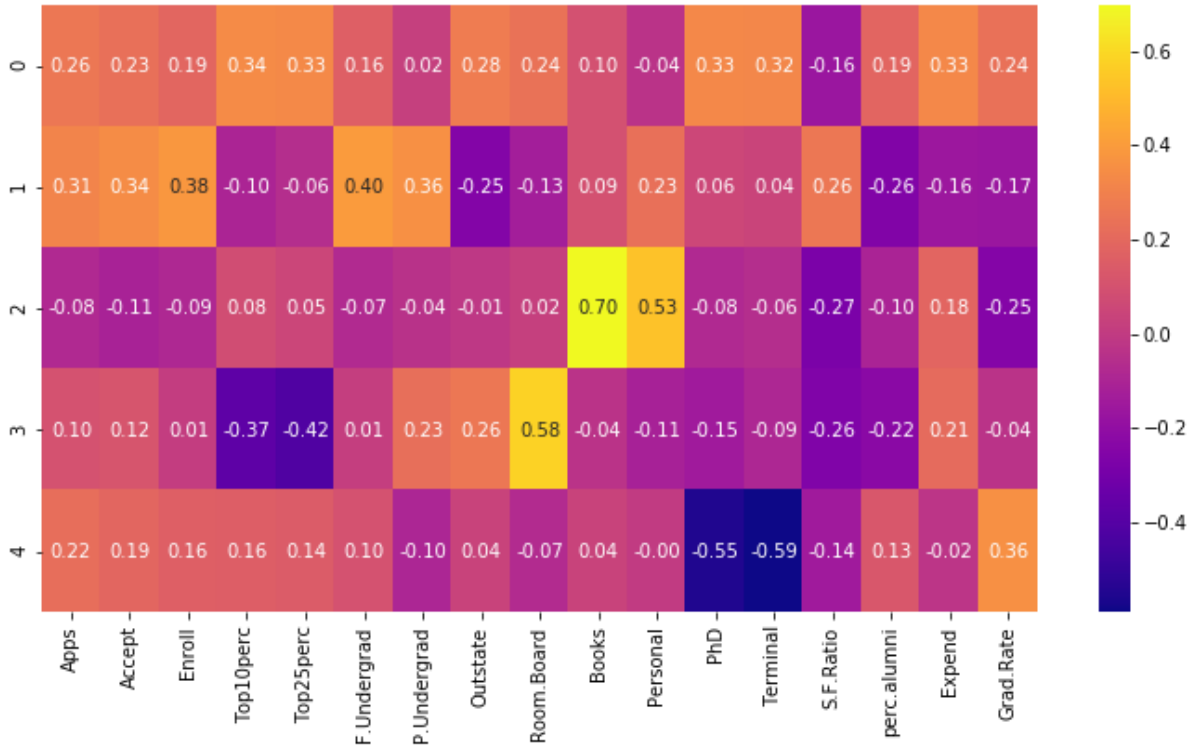
PCA Data frame

	0	1	2	3	4
Apps	0.262172	0.314136	-0.081018	0.098776	0.219898
Accept	0.230562	0.344624	-0.107659	0.118140	0.189635
Enroll	0.189276	0.382813	-0.085530	0.009307	0.162315
Top10perc	0.338875	-0.099319	0.078829	-0.369115	0.157211
Top25perc	0.334691	-0.059506	0.050794	-0.416824	0.144449
F.Undergrad	0.163293	0.398636	-0.073708	0.013950	0.102728
P.Undergrad	0.022480	0.357550	-0.040357	0.225351	-0.095679
Outstate	0.283547	-0.251864	-0.014939	0.262975	0.037275

	0	1	2	3	4
Room.Board	0.244187	-0.131909	0.021138	0.580894	-0.069108
Books	0.096708	0.093974	0.697121	-0.036156	0.035406
Personal	-0.035230	0.232440	0.530973	-0.114983	-0.000475
PhD	0.326411	0.055139	-0.081113	-0.147261	-0.550787
Terminal	0.323116	0.043033	-0.058979	-0.089008	-0.590407
S.F.Ratio	-0.163152	0.259805	-0.274151	-0.259486	-0.142843
perc.alumni	0.186611	-0.257093	-0.103716	-0.223982	0.128216
Expend	0.328956	-0.160009	0.184206	0.213756	-0.022424
Grad.Rate	0.238822	-0.167524	-0.245336	-0.036192	0.356843

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

Answer



Business implication of using the Principal Component Analysis for this case study.

1) This heat map and the colour bar basically represent the correlation between the various feature and the principal component itself.

2) PC1 are more related to Top10perc and Top25perc

3) PC2 are related to F.Undergrad and P.Undergrad.

4) PC3 could be labeled with books.

5) PC4 looks more related to Room.Board

6)From this case study , it is very clear that how Principal component Analysis helped us reduce the dimensionality, From 17 variables that were being used to predict target variable, we are able to extract 6 features that can explain the randomness in the target variable to almost the same degree.