BUSINESS REPORT FOR ADVANCED STATISTICS PROJECT

Table of Contents:

Problem 1 Statement:

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

Exploratory Data Analysis:

	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

The dataset contains 4 variables. “A”, “B” and “Volunteer” is of int64 type and the variable “Relief” is of float64 datatype.

Descriptive Statistics of the Dataset:

	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

Since all the variables is of numerical category, there is no unique values or frequently used categorical variable here.

Even though the variable A and B is numerical, according to the problem statement the only was to use them for proper calculation is to convert them into categorical variables(For Anova calculations).

Check for NULL Values:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   A          36 non-null     category
 1   B          36 non-null     category
 2   Volunteer  36 non-null     int64
 3   Relief     36 non-null     float64
dtypes: category(2), float64(1), int64(1)
memory usage: 984.0 bytes
```

From the above result it is evident that there are no NULL values. The variables 'A' and 'B' is converted into Categorical Variables which is seen above.

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

Null and Alternate Hypothesis for Conducting one-way ANOVA with variable 'A':

Statement Form:

The NULL hypothesis for ANOVA is that the mean is the same for all groups of Column A.

(in other words, the component A poses no significance across groups for providing Relief)

The Alternate hypothesis is that the mean is not the same for all groups of Column A.

(in other words, the component A poses significance across groups for providing Relief)

Statistical Form:

Ho : $\mu1 = \mu2 = \mu3$

Ha : $\mu1 \neq \mu2 \neq \mu3$ or $\mu1 = \mu2 \neq \mu3$ or $\mu1 \neq \mu2 = \mu3$ (at least one of the means is different)

Null and Alternate Hypothesis for Conducting one-way ANOVA with variable 'B':

Statement Form:

The NULL hypothesis for ANOVA is that the mean is the same for all groups of Column B.

(in other words, the component B poses no significance across groups for providing Relief)

The Alternate hypothesis is that the mean is not the same for all groups of Column B.

(in other words, the component A poses significance across groups for providing Relief)

Statistical Form:

Ho : $\mu1 = \mu2 = \mu3$

Ha : $\mu1 \neq \mu2 \neq \mu3$ or $\mu1 = \mu2 \neq \mu3$ or $\mu1 \neq \mu2 = \mu3$ (at least one of the means is different)

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

The NULL and the Alternate Hypothesis is already defined(Please refer 1.1).

Based on the calculation as shown above, we can conclude from the P-Value(4.578242e-07)which shows to be significantly less than the alpha value(Norm : 0.05 and therefore we reject the NULL hypothesis and conclude with Alternate) and therefore to be said that At least one of the means of Variable A is not equal with the other or Variable A proves to be significant in providing Relief .

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.

	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 NULL and the Alternate Hypothesis is already defined(Please refer 1.1).

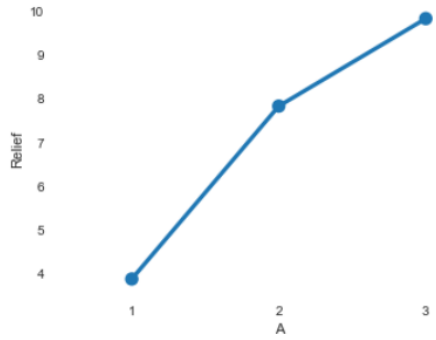
Based on the calculation as shown above, we can conclude from the P-Value(0.00135)which shows to be significantly less than the alpha value(Norm : 0.05 and therefore we reject the NULL hypothesis and conclude with Alternate) and therefore to be said that At least one of the means of Variable B is not equal with the other or Variable B proves to be significant in providing Relief .

PS: The P value for B has significant difference over P value performed on A. The result maybe the same but the effect of A has more significance over effect of B in terms of P value.

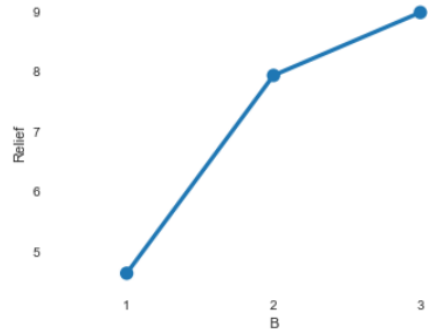
1.4) Analyse the effects of one variable on another with the help of an interaction plot.

What is the interaction between the two treatments?

For showing the interaction of one variable to another a POINTPLOT is used here.



This above graph is nothing but the trend of different groups of variable A in relation to the variable Relief.

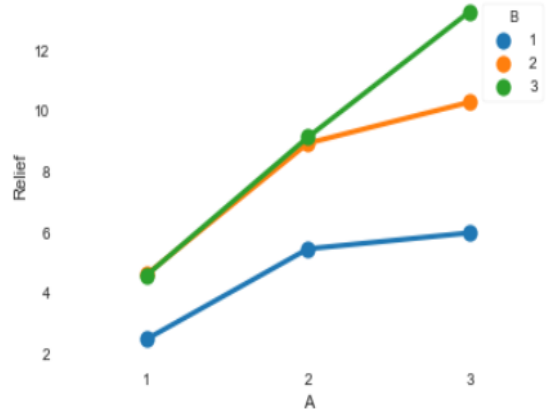


This above graph is nothing but the trend of different groups of variable B in relation to the variable Relief.

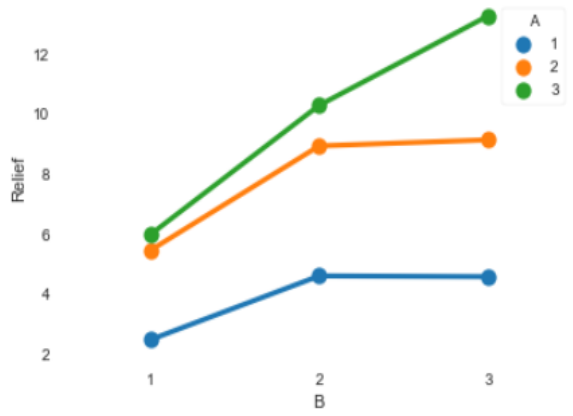
Interaction effect:

It is the simultaneous effect of two or more independent variables on at least one dependent variable as per definition. so here the two ingredients are independent, and the relief is dependent variable. The graph is plotted in such a way that an independent variable will be held as x axis and the dependent variable(Relief) will be a constant on the y axis.

<AxesSubplot:xlabel='A', ylabel='Relief'>



<AxesSubplot:xlabel='B', ylabel='Relief'>



From the above two plots its evident that there is an interaction between the Variables A and B as we can see even an overlap in graph on the left and to further insist that the lines are not parallel(parallel means no interaction).

Therefore, even though we cannot clearly state the proper percentage of the interaction we can say that there indeed interaction between Variable A and B across groups(especially groups 2 and 3).

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 Hypothesis : Means of Variable A and Variable B and the interaction mean(A*B) are equal.

Alternate Hypothesis: At least one of the means from Variable A and Variable B and the interaction mean(A*B) are not equal.

	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

From the above calculations it is evident that All(A, B and A*B) have P values less than alpha(0.05) and therefore we can reject the NULL hypothesis to state that Variable A , Variable B and the interaction of A*B poses significance in proving RELIEF.

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

Firstly when we look into the dataset provided about the experiment taken on various components we can see that this dataset proves to be a perfect case on calculating ANOVA test as it is made of different groups spread across various components and therefore the need for the basic T or Z test is none.

The main purpose of this test is to see whether both ingredients at what levels are required for the treatment that provides many hours of Relief for Hay fever and move from experiment stage to the approval stage for treatment of Hay fever.

Looking at the results of the One-way ANOVA test we can see that Both A and B variable being standalone has a significance in providing RELIEF(Although Variable A was more than precise in this case compared to B).

The two-way ANOVA didn't change this result and provided further evidence when presented with Interaction of A and B.

When investigated the interaction plot and we saw that there is indeed overlap and therefore caused us to go for interaction conclusion. But the level of importance of significance was not the intention there. But as a further measure a TUKEYHSD Multicomparison method is introduced and the result are as follows.

For A with Relief:

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj   lower  upper  reject
---------------------------------------------------
     1      2     3.95  0.001  1.7814 6.1186   True
     1      3     5.95  0.001  3.7814 8.1186   True
     2      3      2.0 0.0755 -0.1686 4.1686  False
---------------------------------------------------
```

For B with Relief:

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj   lower  upper  reject
---------------------------------------------------
     1      2      3.3 0.0164  0.5374 6.0626   True
     1      3     4.35 0.0014  1.5874 7.1126   True
     2      3     1.05 0.6164 -1.7126 3.8126  False
---------------------------------------------------
```

From the above two results we can see how different groups are compared with RELIEF.

Both of them in this case point to the fact that Groups 2 and 3 of ingredient A and B comparatively forms a significance upon providing hours of RELIEF.

Problem 1 Summary:

1.1) The theoretical and statistical way was presented for NULL and Alternate Hypothesis on performing one-way ANOVA with variable RELIEF.

1.2) At least one of the means of Variable A is not equal with the other or Variable A proves to be significant in providing Relief .

1.3) At least one of the means of Variable B is not equal with the other or Variable B proves to be significant in providing Relief .

1.4) Even though we cannot clearly state the proper percentage of the interaction we can say that there indeed interaction between Variable A and B across groups.

1.5) We can reject the NULL hypothesis to state that Variable A , Variable B and the interaction of A*B poses significance in proving RELIEF.

1.6) Business Implication were told by looking into the result of various ANOVA tests performed with the result told that all the Variables do make a significance for providing RELIEF.

Conclusion:

Looking into the data of “Fever.csv”, we saw some insights on how various levels from ingredients A and B poses significance against the providing RELIEF which helps us to decide on whether to go the treatment from the experiment stage to further move across to making it as approved treatment.

Problem 2 Statement:

The dataset Education - Post 12th Standard.csv is a dataset that contains the names of various colleges. This 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.

Basic information about the Dataset:

	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
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

There are totally 18 variables in this dataset.

Categorical Variable : 1 – 'Name'

Numerical Variable : 17 - 'Apps','Accept','Enroll','Top10perc','Top25perc','F.Undergrad','P.Undergrad','Outstate','Room.Board','Books','Personal','PhD','Terminal','S.F.Ratio','perc.alumni','Expend','Grad.Rate'

All of the Exploratory Data Analysis will be documented on the question 2.1.

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

Univariate Analysis(EDA):

NULL check:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   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)
memory usage: 109.4+ KB
```

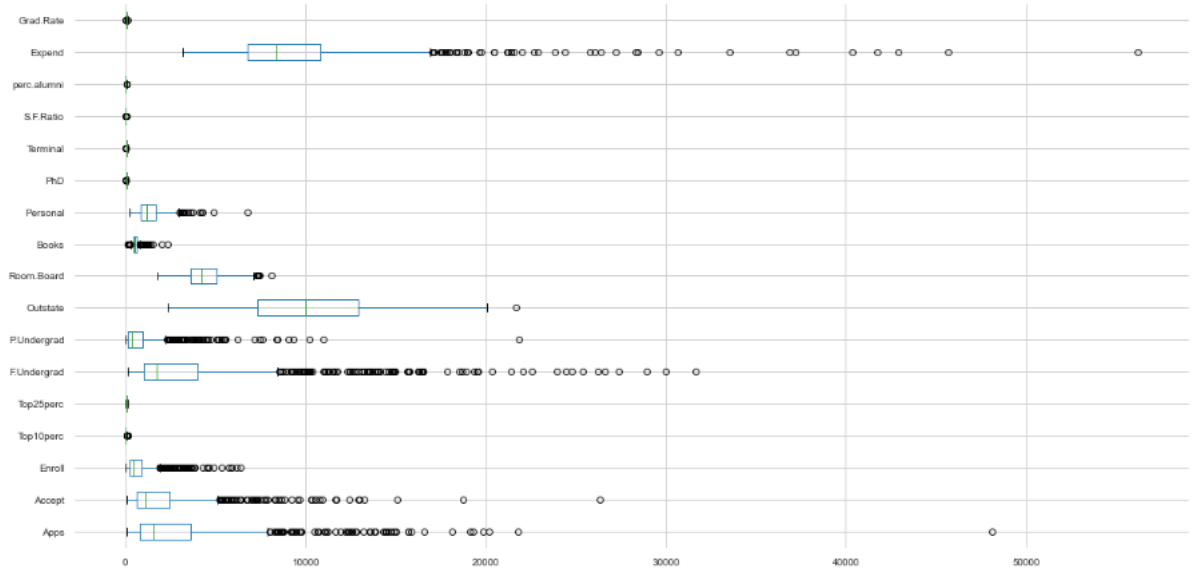
We can see that no NULL values are found in the dataset and therefore we can proceed further in analysing the dataset.

Duplicates Check:

The calculation part is done in the notebook attached and we can see that there is no Duplicates present in the dataset.

Outlier Check:

Usually one of the main methods to view the impact of outliers is to see them through BOXPLOTS.



From the above display on the boxplots for all the variables(Except “Names” which poses no significance in here) we can clearly see that almost all the variables pose some kind of outliers in the data(Except Top25perc). Therefore, highly recommended to perform outlier treatment for this case.

Skewness Check:

```
Apps           3.723750
Accept         3.417727
Enroll         2.690465
Top10perc      1.413217
Top25perc      0.259340
F.Undergrad    2.610458
P.Undergrad    5.692353
Outstate       0.509278
Room.Board     0.477356
Books          3.485025
Personal       1.742497
PhD           -0.768170
Terminal      -0.816542
S.F.Ratio      0.667435
perc.alumni    0.606891
Expend         3.459322
Grad.Rate     -0.113777
dtype: float64
```

From the above result we can see that most of the variables have positive skewness(right skewed data). A very few like ‘PhD’ , ‘Terminal’ and ‘Grad Rate’ shows negative skewness(left skewed data). The distribution based on the skewness report and the boxplot report shows that the data may not be highly normal data.

Data Description(Holds 5-point summary):

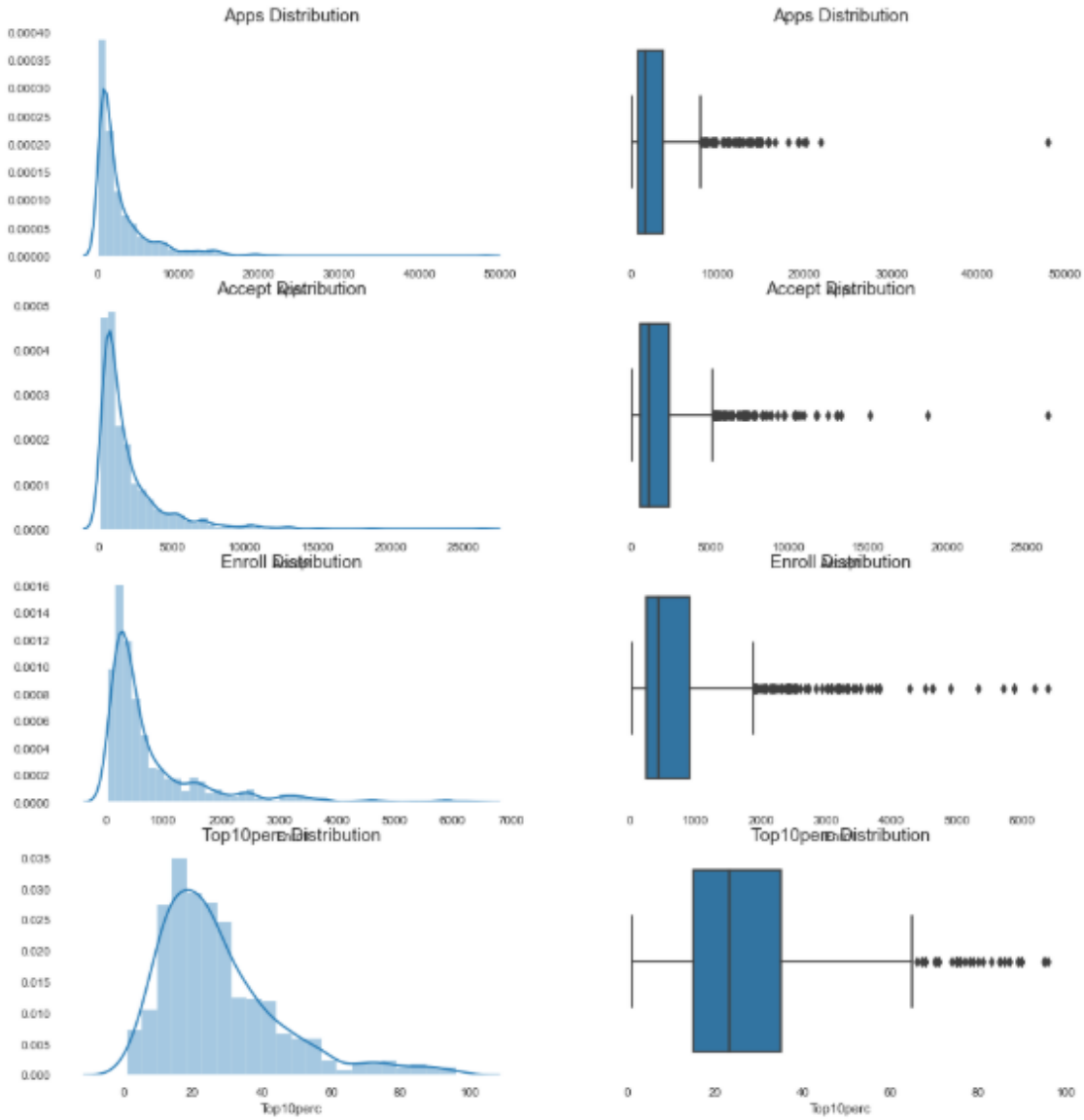
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000

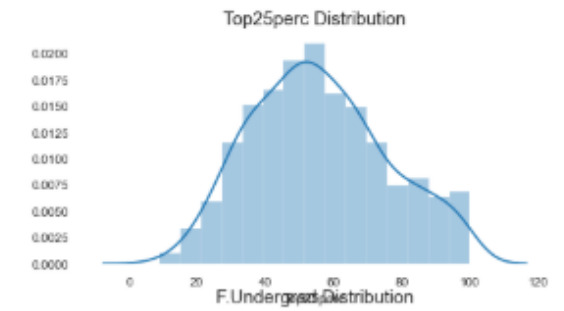
Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00000
1340.642214	72.660232	79.702703	14.089704	22.743887	9660.171171	65.46332
677.071454	16.328155	14.722359	3.958349	12.391801	5221.768440	17.17771
250.000000	8.000000	24.000000	2.500000	0.000000	3186.000000	10.00000
850.000000	62.000000	71.000000	11.500000	13.000000	6751.000000	53.00000
1200.000000	75.000000	82.000000	13.600000	21.000000	8377.000000	65.00000
1700.000000	85.000000	92.000000	16.500000	31.000000	10830.000000	78.00000
6800.000000	103.000000	100.000000	39.800000	64.000000	56233.000000	118.00000

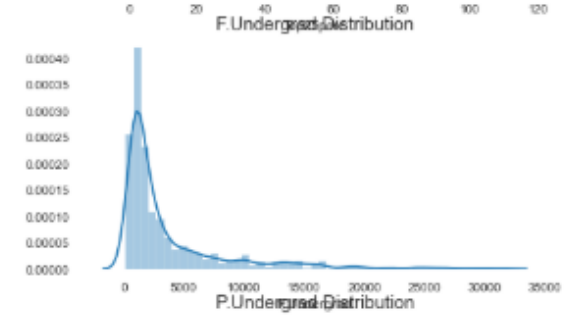
The describe function(describe()) provides us with the information about how much the data is spread across along with the information on the mean, standard deviation, count etc.

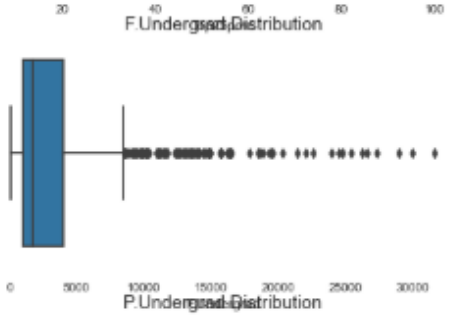
From the above result we can see that data provided is on various scales. The interesting part from the above result is that for some Variables the Standard deviation is more than that of the mean(like Apps, Accept, Enroll and F.Undergrad) which suggests that these variables may not be normally distributed.

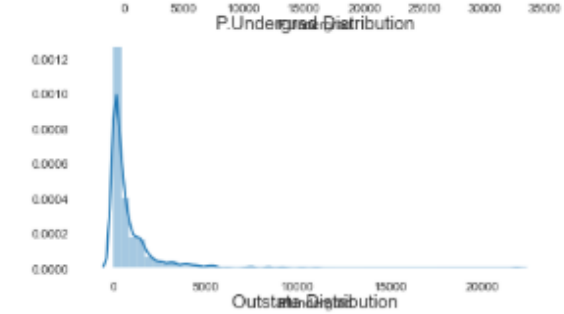
Dist-plot and Boxplot analysis:

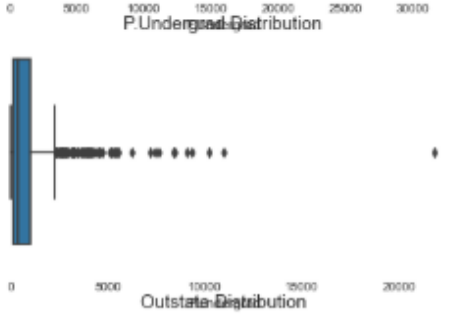


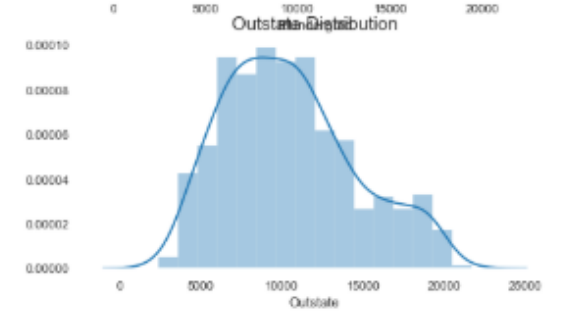


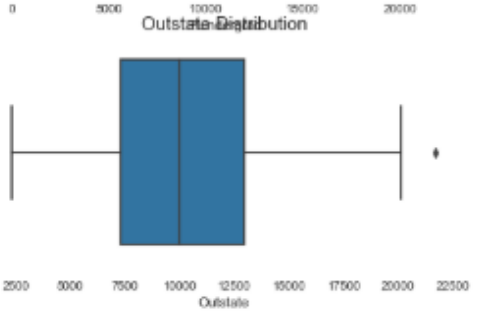


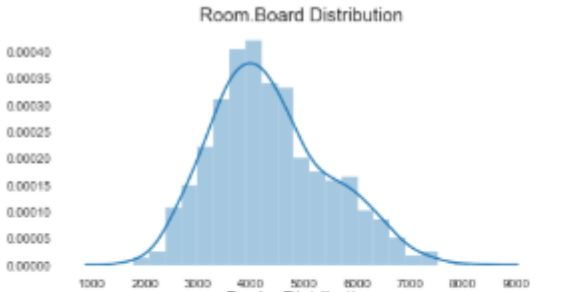


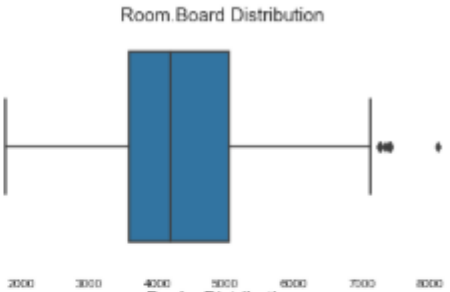


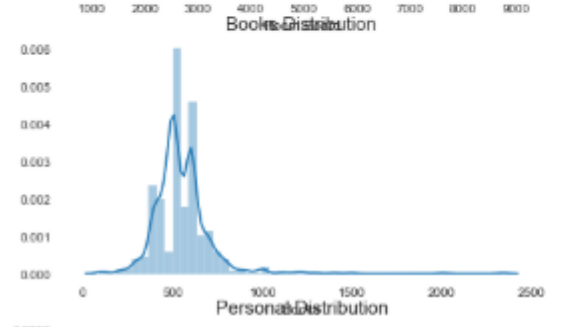


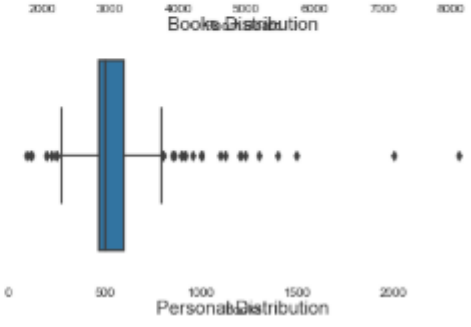


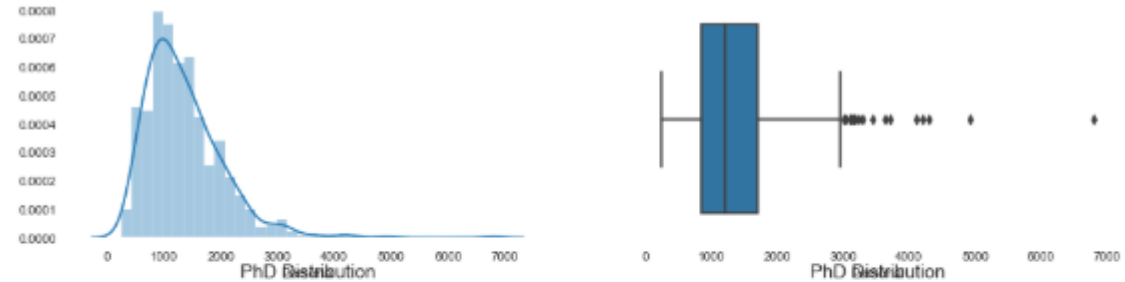


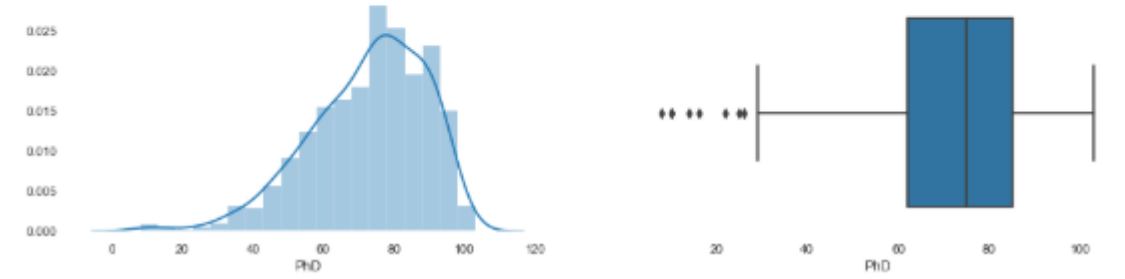


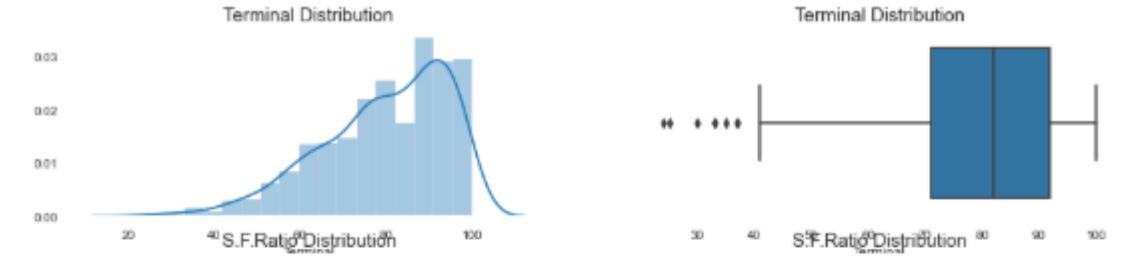


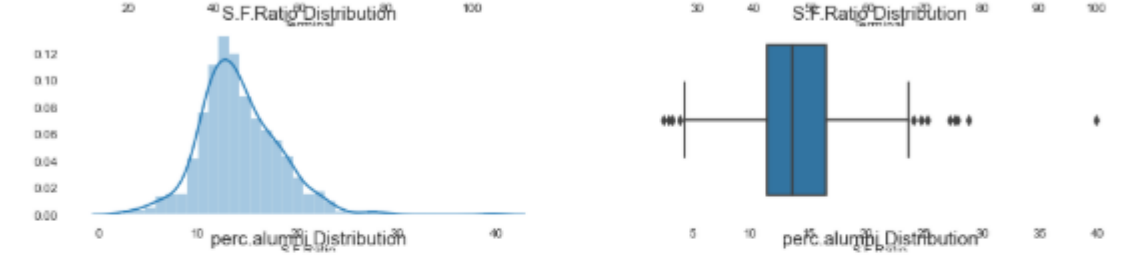


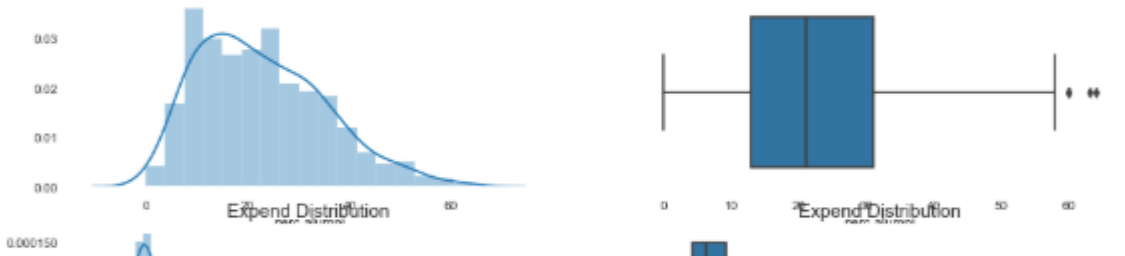


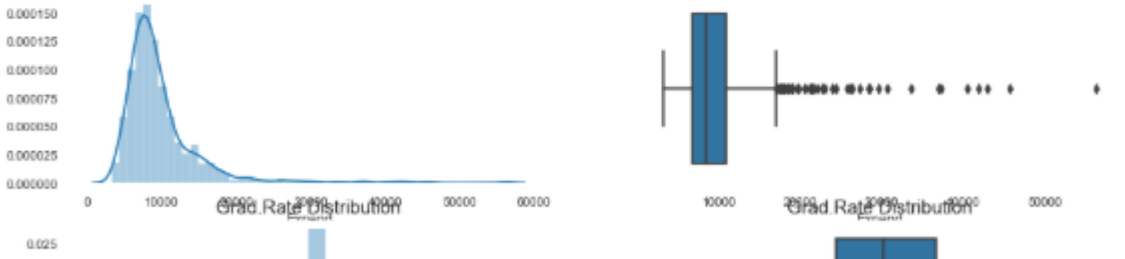


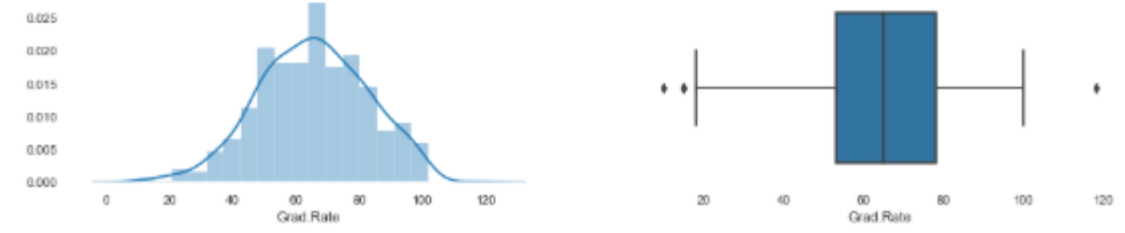










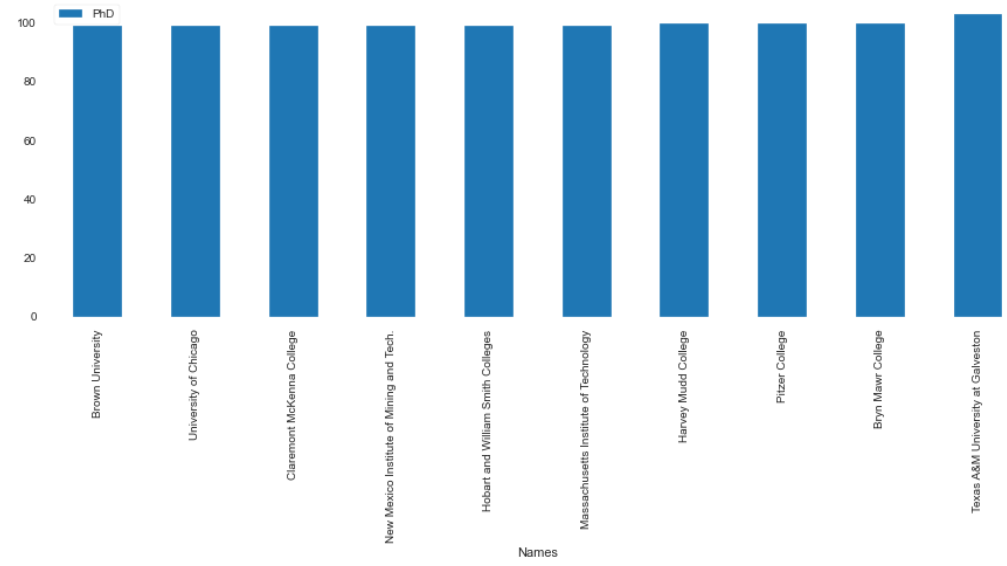


The above results provide a closer insight on the variables to have a much cleaner look on how the data is distributed. It clearly provides a closer look on the results as mentioned earlier for Boxplots as well as the skewness.

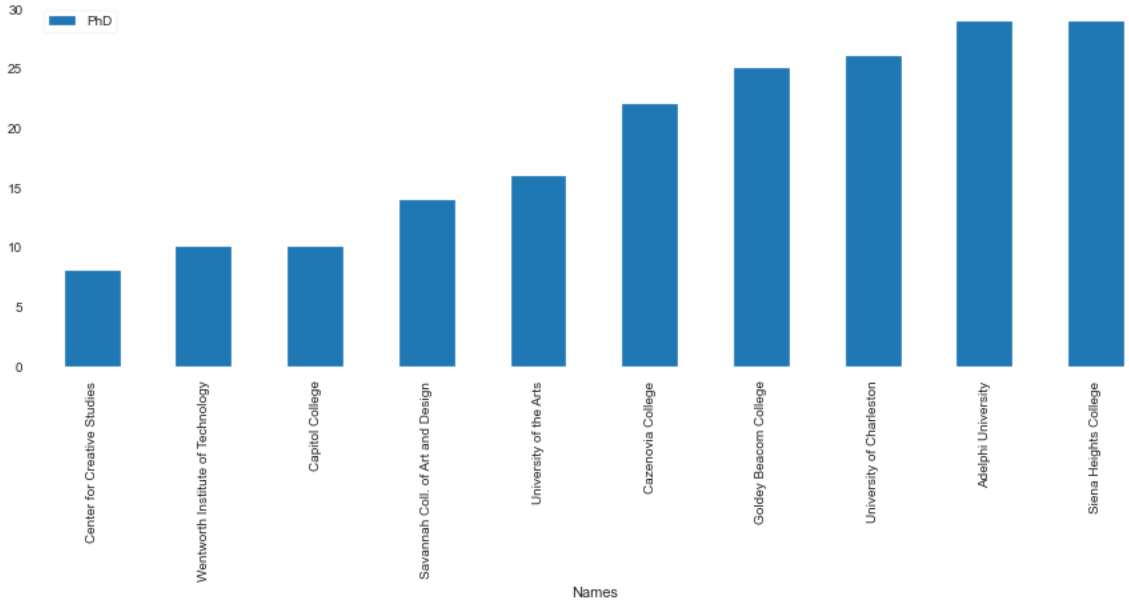
Bivariate Analysis(EDA):

When it comes to performing Bivariate analysis there is a vast analysis that can be performed as the variables is of large number.

But as an sample, here is a boxplot based on Names of the Universities along with the PhD's the faculties in the universities hold.



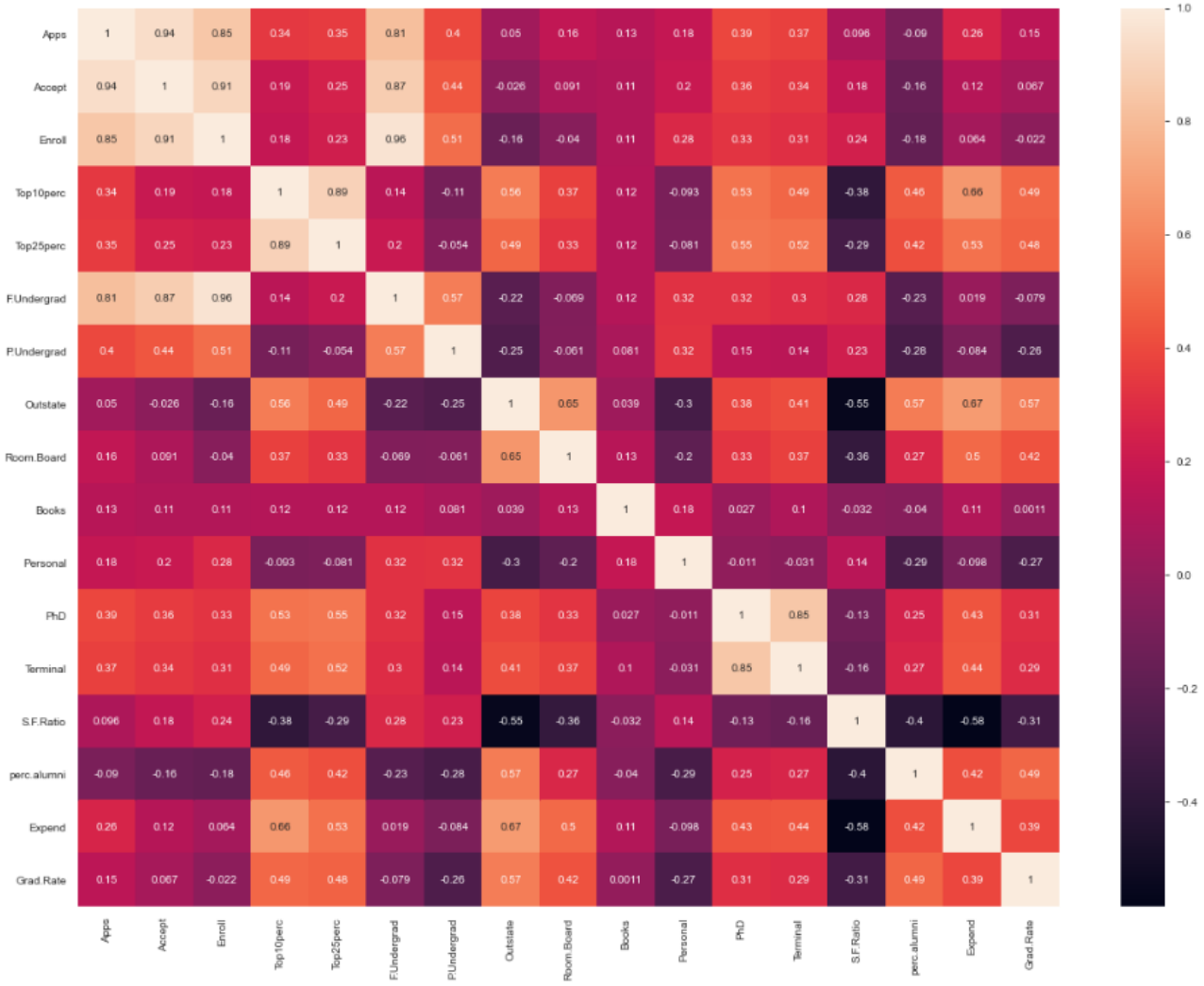
From the above result we can see that the universities which holds the faculties with maximum PhD's belong to 'Texas A&M University at Galveston'.



Similarly, we can do the same for the exact opposite and from the above result we can see that ‘Center for Creative Studies’ holds the faculties with least PhD’s.

Correlation Plot:

Another most important plot that we must look in the case of analysis is the Correlation plot. Below is the Heatmap generated from such Correlation plot of the dataset.

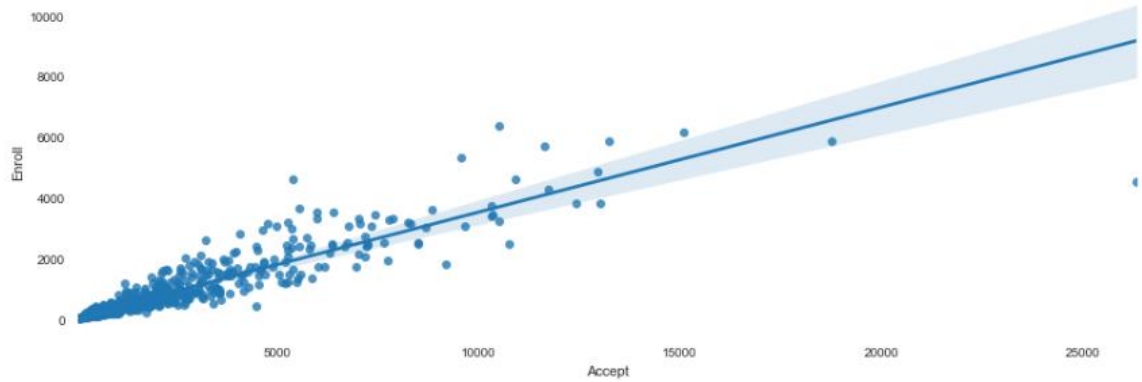


The following conclusions(sample) can be inferred from the heatmap generated.

- Apps, Accept, Enroll and F.Undergrad are highly correlated with each other.
- PhD and Terminal are highly positively correlated(which can prove that PhD’s are also working to move to the most advanced Terminal Degrees is done proper analysis).
- The highest negative correlation belongs to S.F.Ratio and Expend(-0.58) which should also be taken to consideration.

Regplot:

A Regplot was also done for this dataset(one sample case) for Accept and Enroll to see is there a positive correlation to number of applications that were accepted to number of new students enrolled to the program.



From the above plot we can see that there is a high positive regression showed when it comes to variables “Accept” and “Enroll”.

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

Scaling is usual performed as a method of converting the variables of different scales or unit of measures to a defined linear and single unit format. The scaling of variables is one of the most important factors when dealing with PCA.

There are many methods of scaling used for PCA calculations with the most common one’s being

- Standard Scalar(or Z score converter)
- MinMax Scalar
- Logarithmic Transformation
- Exponential Transformation

The most preferred one is Standard Scalar which is generally considered as Industry Standard.

Standard scalar is preferred here over the next commonly used MinMax here even though the data is not gaussian(where minmax is commonly used) is because of the range. The MinMax shrinks the data to a much shorter region even though it might help you get rid of outliers to a small extent.

keeping range into consideration the Standard Scalar is considered here. The main highlight of Standard scalar is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. In other words, the Standard Scalar moves the data towards Unit Variance.

And thus, the Standard Scalar is used as a scaling method here.

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

Covariance is commonly used to determine linear relationship between two variables whereas correlation matrix is generally used to determine strength and direction of the linear relationship between two variables.

The range generally differ when it comes to covariance and correlation where Correlation range between -1 to 1 whereas for the covariance its from negative infinity to positive infinity.

Usually the covariance is preferred over the covariance matrix as it is not affected by scaling, dimensions etc.

Since the range is between -1 and 1, we can easily figure out the linear relationship between two variables.

But these differences wont matter much after scaling is performed as scaling brings them both to an identical standpoint i.e., after scaling both the matrices look almost identical as there are no movement happening after scaling.

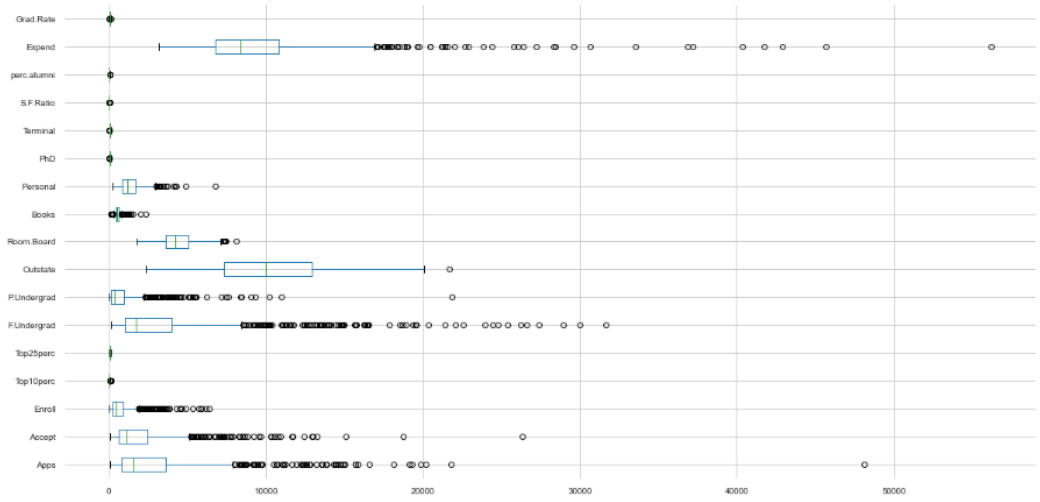
Or in simple terms, after scaling

$$Cov(X,Y) = Corr(X,Y)$$

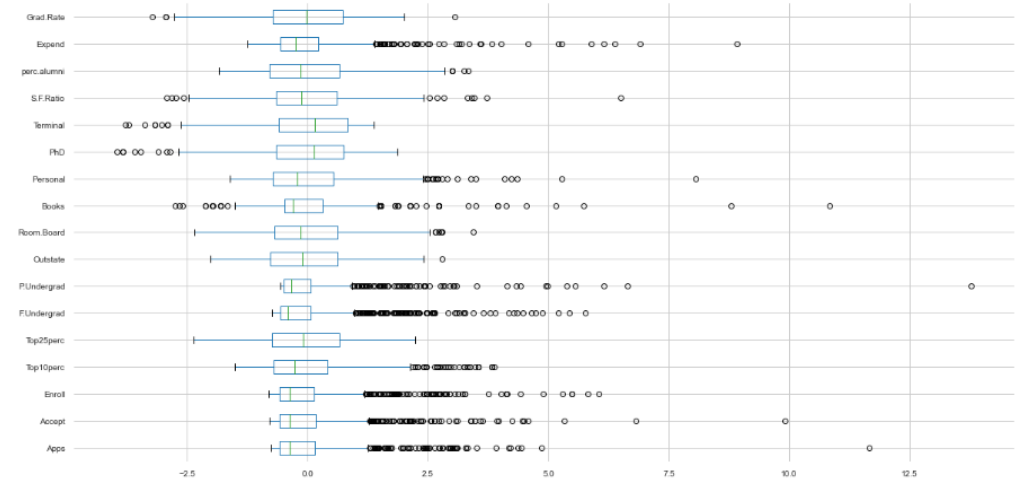
The use case that we have is no exception as after scaling is performed the matrices of both correlation and covariance looked almost identical.

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

Before Scaling:



After Scaling:



Like mentioned before the method to properly visualize the data is by doing BOXPLOT to check for outliers.

From the above two results we can see that in the before scaling plot the data looks highly skewed with and outliers is present almost in all the variables.

When we look at the After-scaling plot, we can see that data looks much more clean, compact and complaint towards calculations. But we can see that even though the distribution looks much cleaner and skewness looks better the outliers are still significant(as now the distribution comes of range -3 to +3 excluding the outliers). They have been reduced a bit but doesn't look highly impactful on outliers.

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

The below built values is based on the Standard Scaled Data frame without treating outliers(As it is not necessary)

Covariance Matrix:

```
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   -0.08891441  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 -0.08891441  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]]
```

Eigen Values:

```
array([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, 0.00802464, 0.1439785 ,
       0.16779415, 0.22861096])
```

Eigen Vectors:

```
array([[-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]])
```

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

<u>Based on Eigen Vectors:</u>

PC0 = (-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)

Or (Direction change is observed which may be the result of libraries used, but the contribution hold true irrespective of direction)

Based on Dataframe from calculating PCA components(From 2.7):

PC0 = (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)

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.

Note : Standard Scaler is performed for the Data for further calculations

Cumulative Distribution of Eigen values:

```
[32.02062819886915,
 26.34021443611248,
 6.9009165542224995,
 5.922989222926291,
 5.488405110358482,
 4.98470095455744,
 3.5588714917466553,
 3.4536213369992645,
 3.1172336798217217,
 2.3751915258937992,
 1.8414263209386887,
 1.296041400123535,
 0.9857541228001165,
 0.845842335083003,
 0.517125583373192,
 0.2157540100727578,
 0.13528371610095175]
```

Cumulative values of the eigenvalues:

```
array([ 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.        ])
```

From the values we can see the increase in percentage of variation leading to a complete variation of 100%. The values are used here such that the variation of second component is the cumulative addition of variation of fist and the second component(Example for second component we add 32.0206282 and 26.340214436 to get 58.36084263).

Deciding on the optimum number of principal components:

There are three commonly used methods to decide on the number of principal components. These are just guidelines as there are not exact method to do this.

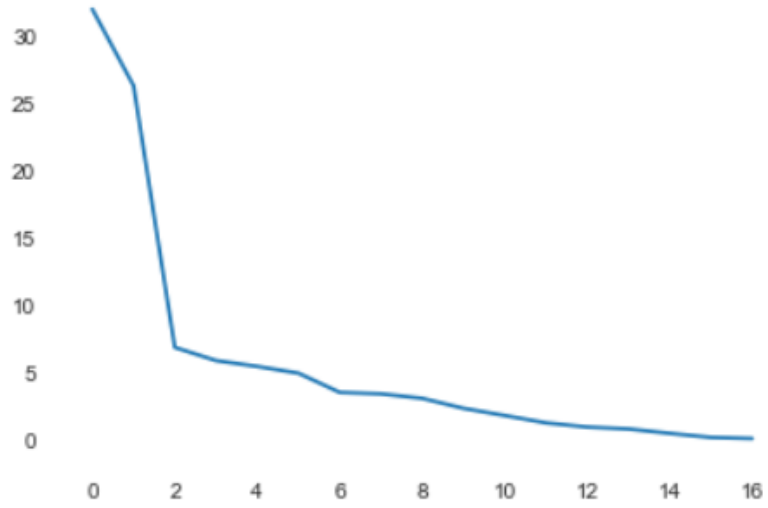
They are(from the order of most to least accurate):

- Cumulative variances pose at least more than 70 percent(better to go for minimum 80 percent) of variance.
- Eigen values: components with values more than 1(not compulsory but recommended).
- Scree plot or elbow plot : The scree plot will show a steep drop in variance explained with increase in number of PC's. The part where the elbow breaks with the near parallel increase will be taken as the number of optimum components.

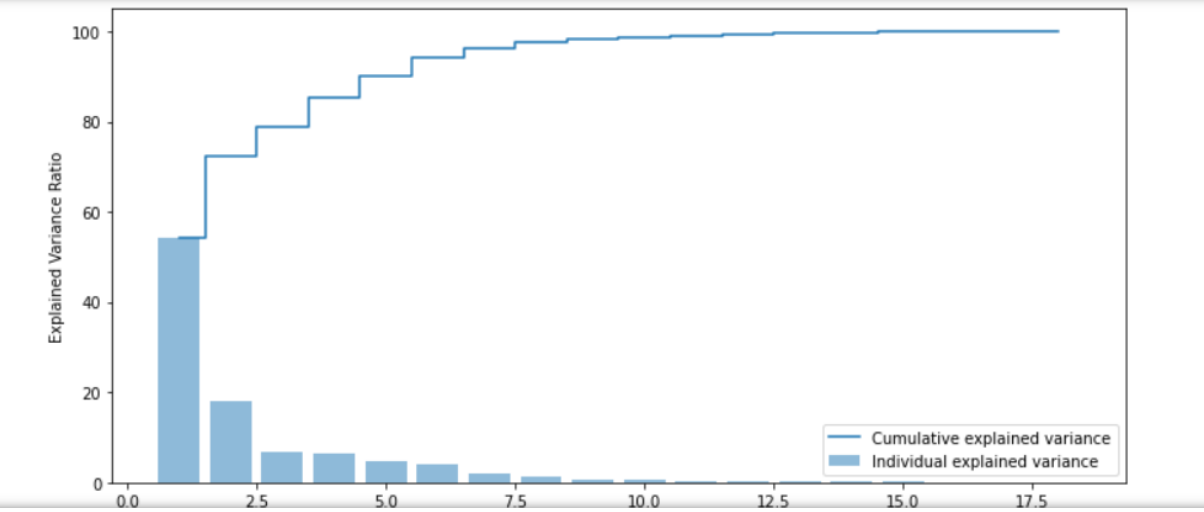
From the calculation on the Scaled Dataframe we can see that more than 80 percent variation is attained for 6 PCA's.

The number of eigen values taken for this is 6(even though last two are not more than 1, it is near 1 and provides a better show on variation).

Scree Plot:



The scree plot shows the same result as the other two as we can see that the line levels out with very minimum variation after the 6th component.



The above graph shows the clear indication of the individual and cumulative explained variance and how it spreads.

Eigen Vectors indication:

Eigen vectors usually indicate the contribution of each individual variable in terms of actual percentage and direction to a component. Eigen values and Eigen vectors of a covariance/correlation(as they are same on a scaled DF) helps in identifying the direction and magnitude of feature space. An eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it.

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

The PCA is done through a package called “sklearn”.

We have used the number of components(nComponents = 6) as 6.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0.641055	-0.331398	0.091256

Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0.317056	-0.176958	0.205082	0.318909	0.252316
0.046429	0.246665	-0.246595	-0.131690	-0.169241
-0.066038	-0.289848	-0.146989	0.226744	-0.208065
-0.519443	-0.161189	0.017314	0.079273	0.269129
0.204720	-0.079388	-0.216297	0.075958	-0.109268
0.154928	0.487046	-0.047340	-0.298119	0.216163

The components are converted to a Data frame and exported as a CSV file attached along with the notebook(all the indexes from 0 to 5 indicates the PCA components)

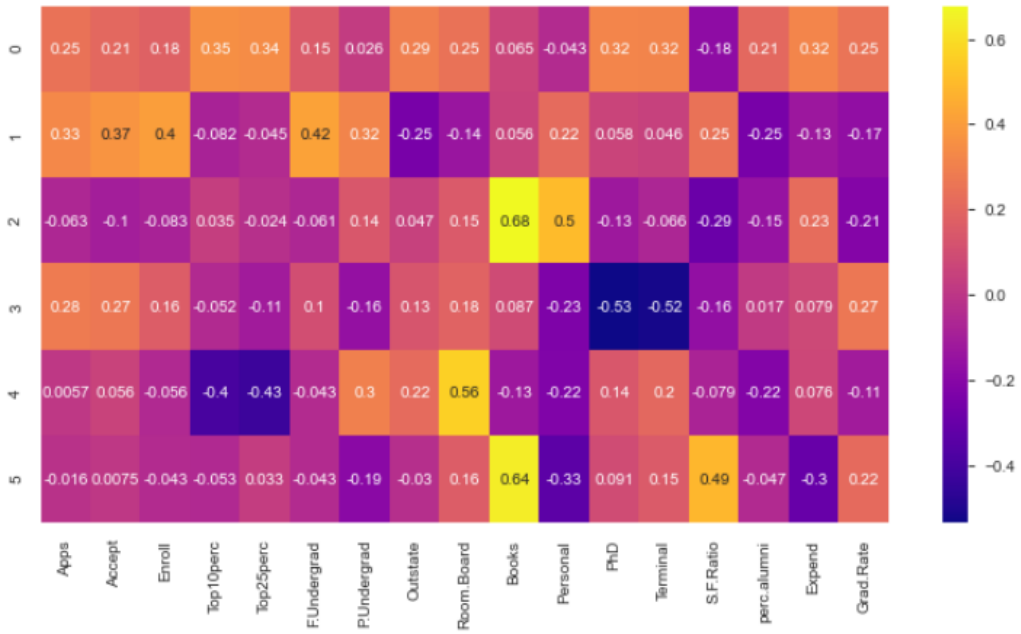
2.8) Mention the business implication of using the Principal Component Analysis for this case study.

Principal Component Analysis is basically a Dimensional Reduction technique used when there are more variables highly correlated with each other. It is used to reduce the correlated variables to a selective number of components without compromising any data loss.

PCA is used in EDA for making the necessary arrangements to avoid multicollinearity and preparing for MODEL creation.

It is done by following two steps,

- Building a covariance(or correlated) matrix on scaled data.
- Performing eigenvalue decomposition on the covariance matrix.



By knowing about the basic information about the PCA we can see that the use for PCA is at large here as the variables are in huge numbers with high correlation. Therefore, it is required to reduce to a minimum number of components with losing any information and avoiding multicollinearity. We can look at the above PCA heatmap to see the number of components and their contributions for each of the variables. An example we can use here is to look at the components 3 which holds the maximum variation for Books(0.68 or 68 percent) and therefore we can use and rename the PCA3 as PCA_Books to better identify the use of this component.

It is always better and advisable to remove redundant predictors when it comes to building a machine learning model(in some cases may not be required).

Problem 2 Summary:

2.1) All the basic EDA along with the univariate and the Bivariate analysis were performed and analysed.

2.2) Standard Scaler is used here and the reason for that have been mentioned.

2.3) Covariance and Correlation matrix has been compared along with their presence after scaled data.

2.4) Outlier prediction was performed through BOXPLOTS before and after scaling and their inferences were explained.

2.5) Covariance matrix was built and Eigen values and Eigen vectors were calculated.

2.6) Explicit form of the first PC was given as a proper equation.

2.7) PCA was performed on the scaled data and the results were explained.

2.8) Business implications on performed PCA were explained.

Conclusion:

Looking into the data of “Education - Post 12th Standard.csv”, we saw some interesting insights on such varied variables and how each of these variables interact with each other. The combinations possible EDA to be performed on this dataset is huge, but some of these have been mentioned in the document. When it comes to the PCA analysis the scaling function mentioned here is Standard Scalar, but in the notebook, even MINMAX and Log transformations are used as a practise and the results have been published.