

Statistics

“ADVANCED STATISTICS”

STATISTICS IN DATASCIENCE :

Statistics is used to process complex problems in the real world so that Data Scientists and Analysts can look for meaningful trends and changes. Data Science includes techniques and theories extracted from statistics, computer science, and machine learning.

BUSINESS REPORT ON ADVANCED STATISTICS

Course Name: PGP-DSBA (Online)

Module Name: Advanced Statistics

Submitted by: Shubhank katarey

Submission Date: 16 May 2021

TABLE OF CONTENTS:

Problem 1A

- ⊕ Problem Statement
- ⊕ Summary of the problem
- ⊕ Loading the data
- ⊕ Head of the dataset
- ⊕ Tail of the dataset
- ⊕ Five number summary
- ⊕ Information about the data
- ⊕ State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.
- ⊕ Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
- ⊕ Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Problem 1B:

- ⊕ What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the 'point plot' function from the 'seaborn' function]
- ⊕ Analyzing the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

- ⊕ Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

Problem 2

- ⊕ Problem Statement
- ⊕ Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
 - ⊕ Head of the data
 - ⊕ Info about the dataset
 - ⊕ Describing the data/Five number summary
 - ⊕ Checking the null values
 - ⊕ Checking the zeros
 - ⊕ Numerical data
- ⊕ Distribution and the Boxplots
- ⊕ Pair plot
- ⊕ Correlation Matrix
- ⊕ Scaled Data
 - ⊕ Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
- ⊕ Covariance matrix
- ⊕ Correlation matrix
- ⊕ Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]
 - ⊕ outliers in the dataset before scaling
 - ⊕ outliers in the dataset after scaling

- ⊕ Extract the eigenvalues and Eigen vectors [print both]
 - ⊕ Eigen values
 - ⊕ Eigen vectors
- ⊕ Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features.
- ⊕ Bartlett's Test of Sphericity:
- ⊕ Performing KMO Test
- ⊕ Percentage variance explained by Eigen values
- ⊕ Cumulative values explained by Eigen values
- ⊕ Scree Plot
- ⊕ PCA components
- ⊕ Consider 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?

Problem 1A:

Problem Statement:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

Summary of the Problem:

In the dataset named 'SalaryData', the information is collected of 40 individuals about their educational qualification and occupation. To understand the dependency, their salary information is also noted. The collected data has 3 levels of educations and 4 levels of occupation.

We need to perform hypothesis testing on this dataset or the population and also need to do analysis on the data to get the dependency on dependent variable 'Salary'.

Head of the data:

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

There are 3 columns: Education, Occupation and Salary.

Tail of the data:

	Education	Occupation	Salary
35	Bachelors	Exec-managerial	173935
36	Bachelors	Exec-managerial	212448
37	Bachelors	Exec-managerial	173664
38	Bachelors	Exec-managerial	212760
39	Doctorate	Exec-managerial	212781

Describe the data:

	Salary
count	40.000000
mean	162186.875000
std	64860.407506
min	50103.000000
25%	99897.500000
50%	169100.000000
75%	214440.750000
max	260151.000000

We can see five number summary and various other information about the data like max, min, 75%, 50%, standard deviation...

Information about the data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Education    40 non-null      object  
 1   Occupation   40 non-null      object  
 2   Salary        40 non-null      int64   
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

We can see information like data types, non-null counts, RangeIndex...

State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

Answer:

Null and Alternate hypothesis for Education:

Null Hypothesis (H0): The mean salary of all the 40 individuals is equal at all education level.

Alternate Hypothesis (Ha): The mean salary of all the 40 individuals are different for atleast one kind of education.

Null and Alternate hypothesis for Occupation:

Null Hypothesis (H0): The mean salary of all the 40 individuals is equal for all kind of Occupation.

Alternate Hypothesis (Ha): The mean salary of all the 40 individuals are different for at least one kind of Occupation.

Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

There are different kind of education which is influencing the salary of every individuals.

Variance in the salary caused by different education level.

Difference in salary of few individuals is because of the difference in their education.

So now, we can see that the p-value is less than the significance level(0.05), hence we can reject the null hypothesis and conclude that the mean salary is different for at least one of the individual.

Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

So now, we can see that the p-value(0.45 as above) is greater than the significance level(0.05), hence, in this case we fail to reject the null hypothesis.

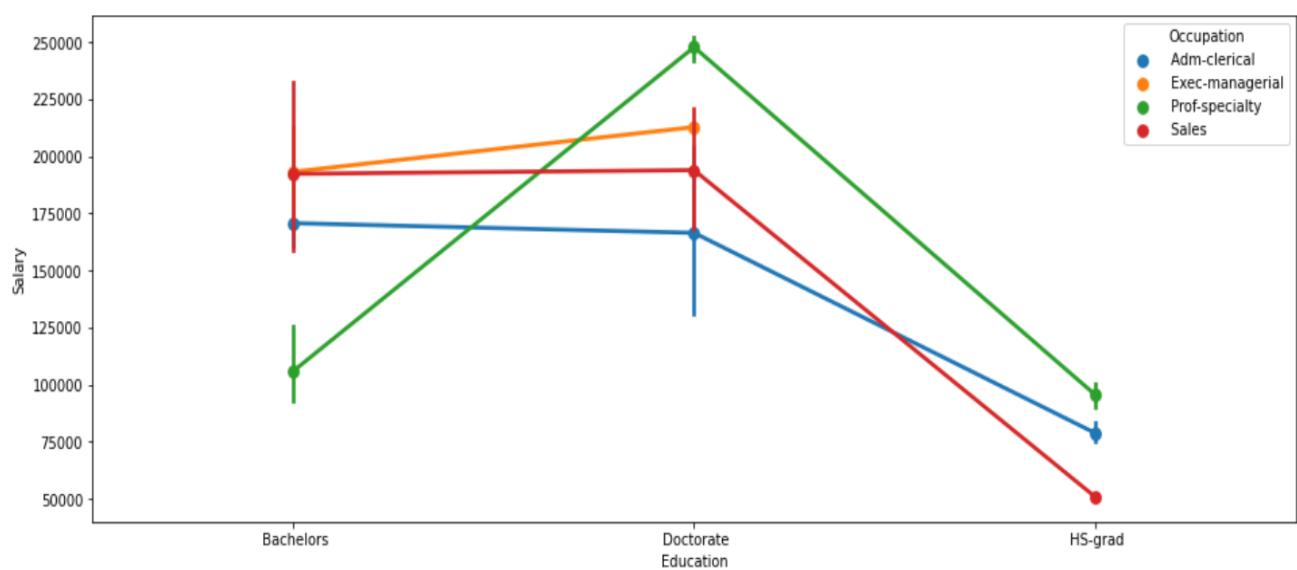
Problem 1B:

Q1. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘point plot’ function from the ‘seaborn’ function]

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

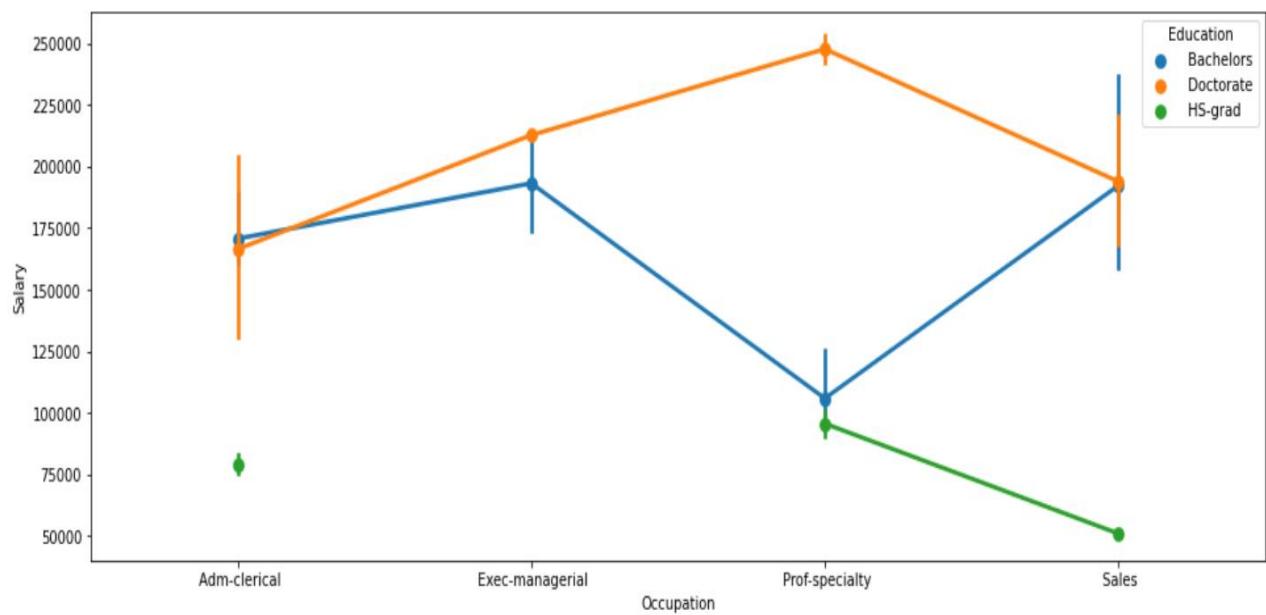
We can see that the p-value in one of the treatments is greater than alpha(0.05).

Analysis of the effect of occupation on education:

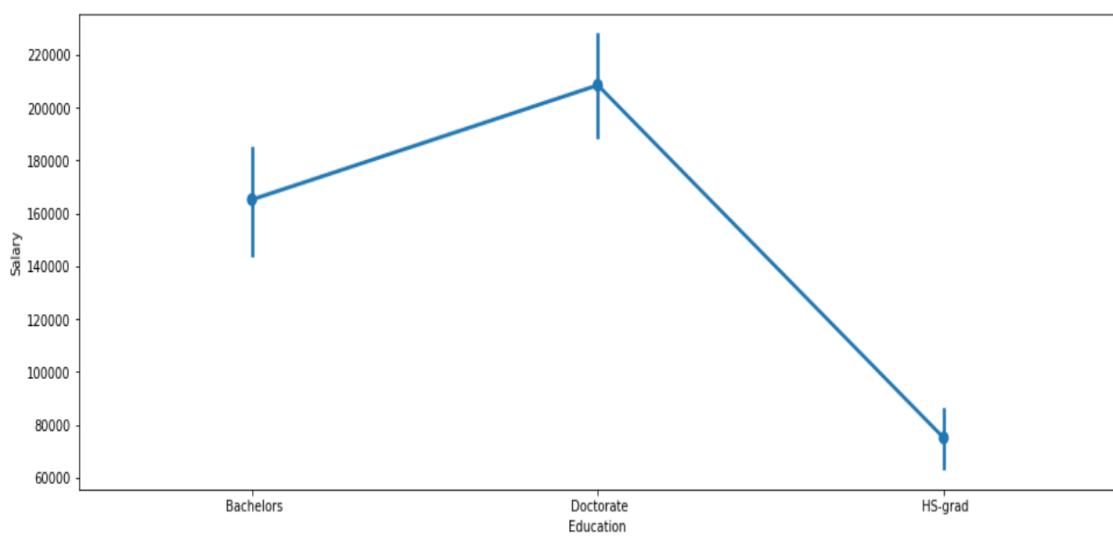


From the above plot we can see that the salary is high of the individuals of doctorate having the occupation of their profession or specialty. And the salary is lowest if the higher secondary educated individual is in the occupation of sales.

Analysis of the effect of education on occupation:



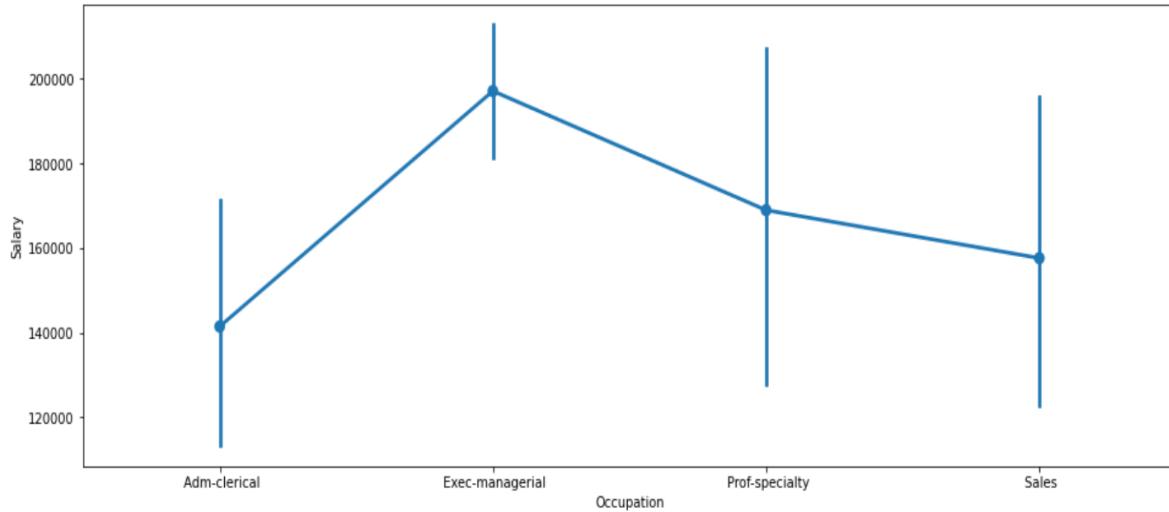
Effect of education on Salary:



Higher secondary graduate has the lesser salary than the doctorate and bachelors.

Doctorate has the highest salary.

Effect of Occupation on Salary:



We can see that exec-managerial occupation has the highest salary.

Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

We can see the little change in p-value of "occupation" without the interaction effect.

Here p-value is less than significance value(0.05) for Education, that means we can reject the null hypothesis.

There is very minor change in p-value of Occupation that is greater than significance value(0.05), so we fail to reject the null hypothesis.

The impact on dependent variable 'Salary' is much due to the 'Education' and the joint interaction effect of 'education' and 'occupation' together.

Explain the business implications of performing ANOVA for this particular case study.

Answer:

The impact on dependent variable 'Salary' is much due to the 'Education' and the joint interaction effect of 'education' and 'occupation' together.

The combined effect of education and occupation together makes a great impact on individuals salary.

Problem 2:

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

Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

Head of the dataset:

2]:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9

Info about the data:

```
<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
```

As we can see that there are 777 non-null rows, 1 column of object data type, 1 column of float data type and 16 columns of integer data type in the dataset.

Description of the dataset:

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.797941	12.338089	1.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Observation:

1. Minimum No. of applications received are 81 and the maximum are 48094
2. Minimum No. of applications accepted are 72 and the maximum are 26330
3. Mean cost for room and board comes out to be 4357 rupees.
4. Maximum cost of the books for a student is 2340 rupees.

Count of 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
```

As we can see that there are no missing values in the dataset.

Count of zeros in the columns:

```
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    2
Expend         0
Grad.Rate      0
dtype: int64
```

As we can see that, only 'perc.alumni' column has 2 zeros.

Dataset of Numerical columns for further analysis:

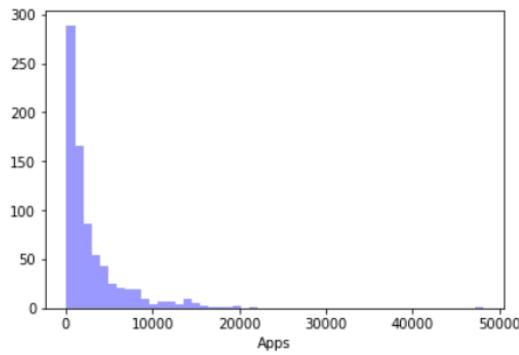
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni
0	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	
1	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	
2	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	
3	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	
4	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	
...
772	2197	1515	543	4	26	3089	2029	6797	3900	500	1200	60	60	21.0	
773	1959	1805	695	24	47	2849	1107	11520	4960	600	1250	73	75	13.3	
774	2097	1915	695	34	61	2793	166	6900	4200	617	781	67	75	14.4	
775	10705	2453	1317	95	99	5217	83	19840	6510	630	2115	96	96	5.8	
776	2989	1855	691	28	63	2988	1726	4990	3560	500	1250	75	75	18.1	

777 rows × 17 columns

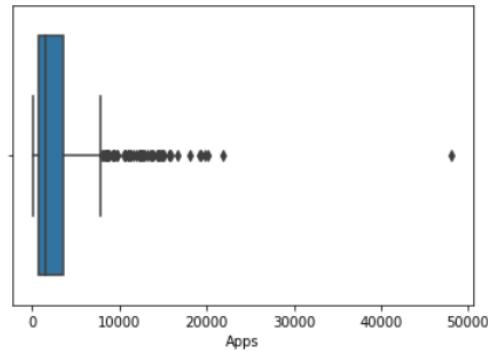
Removed the columns other than numerical (integer and float data types) values for further analysis of the data.

Distribution and Boxplots of the Numerical columns of the dataset:

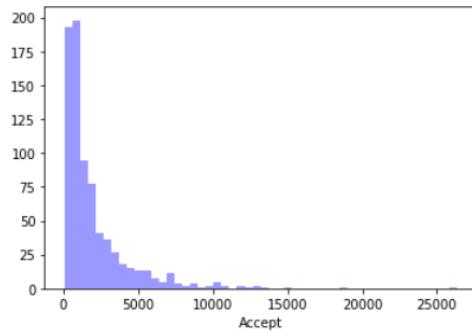
Below is the distribution of :Apps



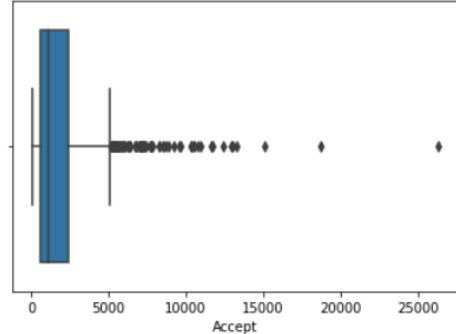
Below is the box plot of :Apps



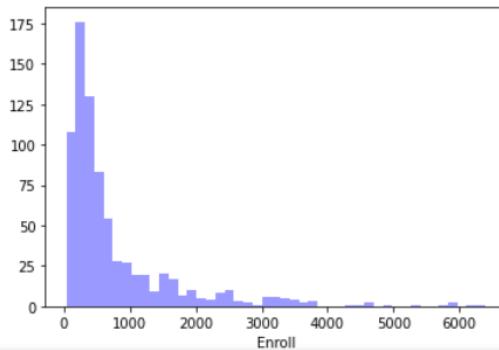
Below is the distribution of :Accept



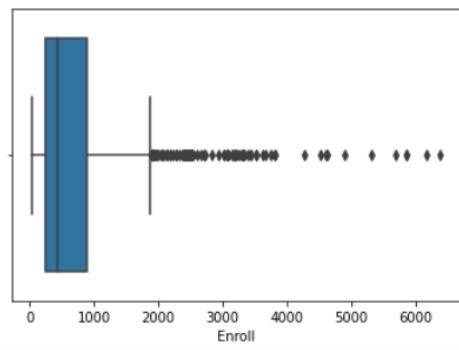
Below is the box plot of :Accept



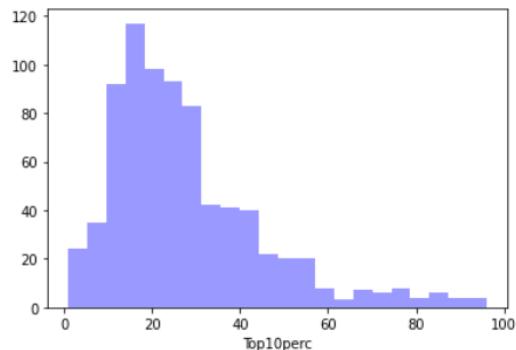
Below is the distribution of :Enroll



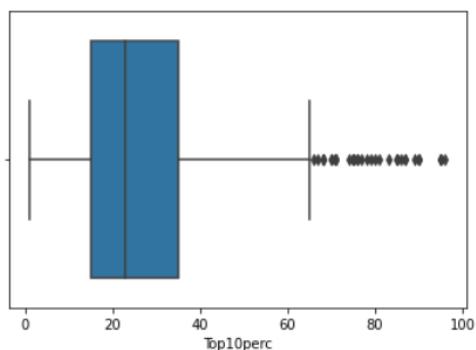
Below is the box plot of :Enroll



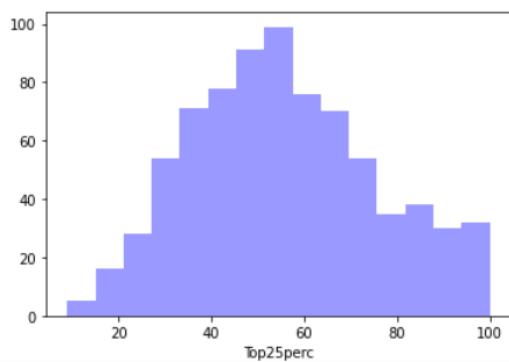
Below is the distribution of :Top10perc



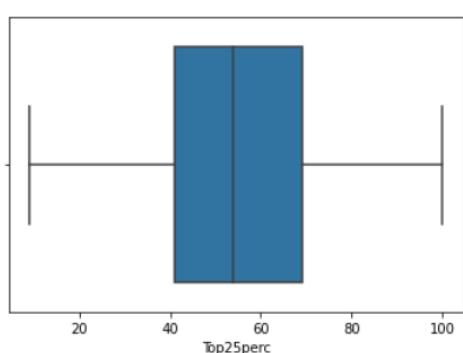
Below is the box plot of :Top10perc



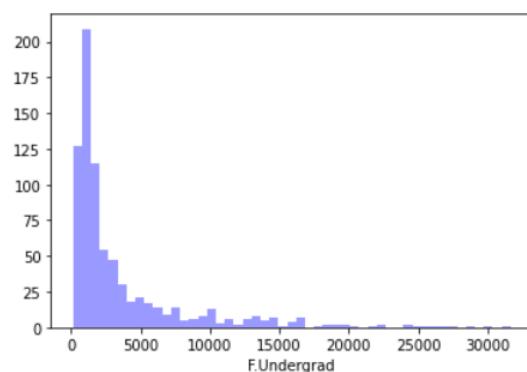
Below is the distribution of :Top25perc



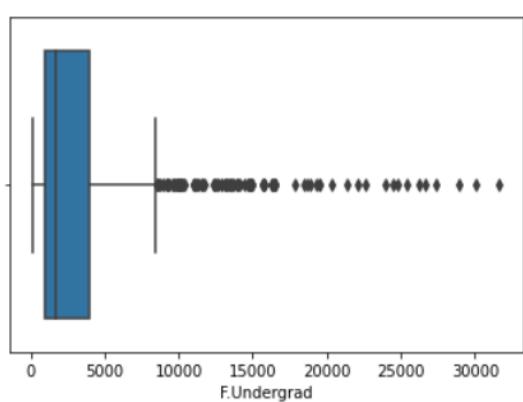
Below is the box plot of :Top25perc



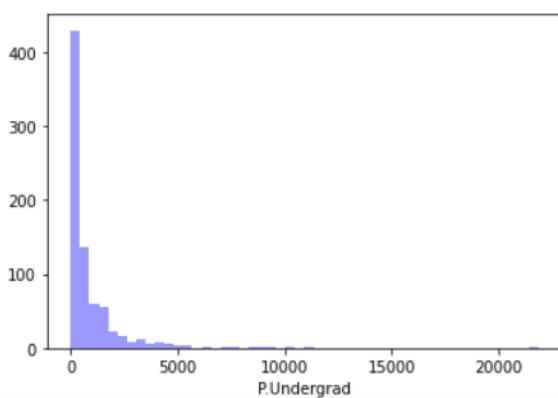
Below is the distribution of :F.Undergrad



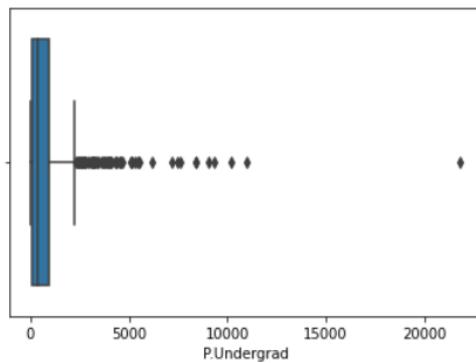
Below is the box plot of :F.Undergrad



Below is the distribution of :P.Undergrad



Below is the box plot of :P.Undergrad



Correlation matrix :

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	1
Apps	1.00000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	C
Accept	0.943451	1.00000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	C
Enroll	0.846822	0.911637	1.00000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	C
Top10perc	0.338834	0.192447	0.181294	1.00000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	C
Top25perc	0.351640	0.247476	0.226745	0.891995	1.00000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	C
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.00000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	C
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.00000	-0.253512	-0.061326	0.081200	0.319882	0.149114	C
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.00000	0.654256	0.038855	-0.299087	0.382982	C
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.00000	0.127963	-0.199428	0.329202	C
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.00000	0.179295	0.026906	C
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.00000	-0.010936	C
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.00000	C
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	C
perc.alumni	-0.091649	-0.161391	-0.181458	0.452853	0.418289	-0.229185	-0.282213	0.565162	0.272085	-0.039118	-0.281762	0.248035	C
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	C
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	C

Scaled Data:

Most of the time, the variables present in the data are of different scales, for example one variable having 4 digits of numeric values and other having single digit. So, it becomes difficult to compare these variables. That's why we use feature scaling to standardize the range of features of data. It is very important step. In this method, we convert different scales of measurement into single scale. We will use zscore to normalize the data (only for numerical data).

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	Expend	Grad.Rate
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.175657	
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176	-0.523535	
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341	-0.523535	
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657	-0.523535	
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535	-0.523535	

Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].

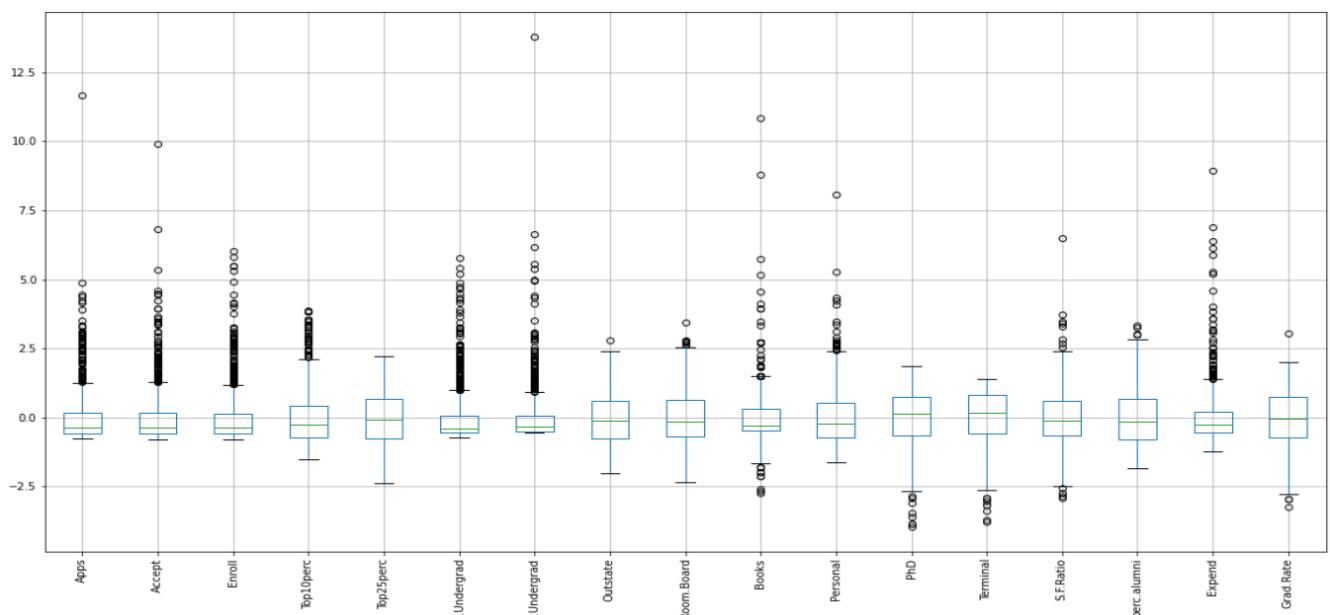
Correlation is a measure used to represent how strongly two random variables are related to each other. It is basically the scaled form of covariance.

Covariance is nothing but a measure of correlation. Covariance indicates the direction of the linear relationship between variables.

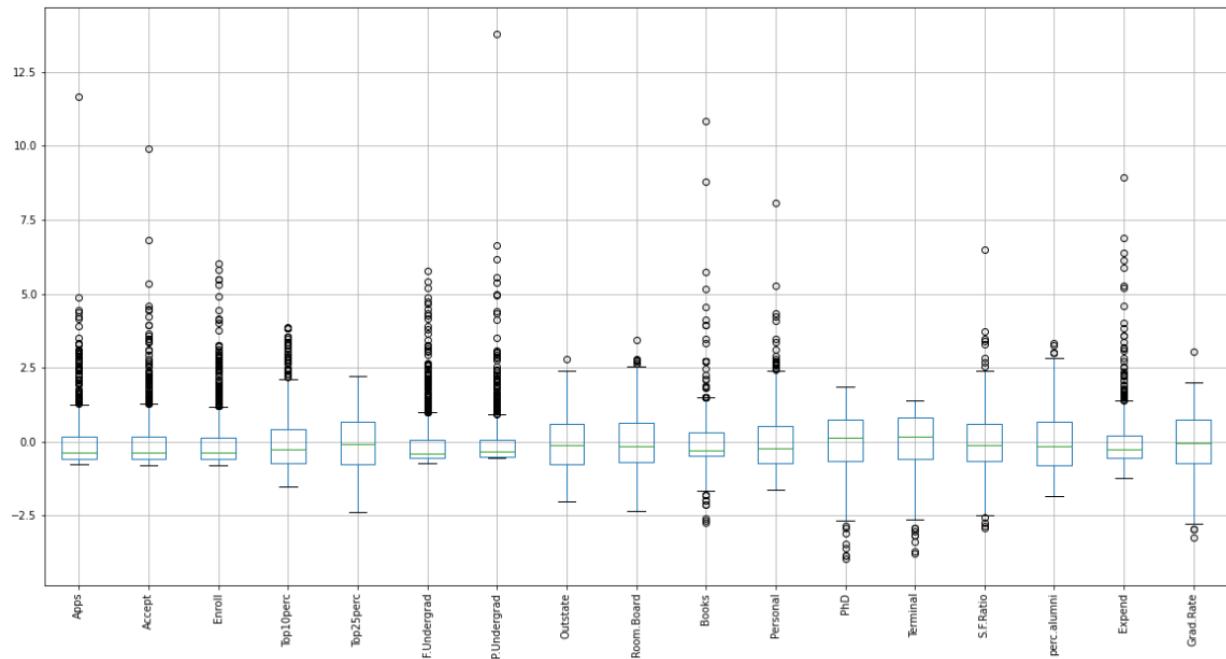
In our below correlation matrices we can see the correlation between all the 17 variables in the dataset but in covariance it is linear relationship between variables.

Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

Outliers before scaling:



Outliers after scaling:



Outliers are present in the data. We can see clearly the difference after the scaling of the data. Now every variable is showing on single scale.

Extract the eigenvalues and eigenvectors. [Print both]

Eigen Values

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

Eigen Vectors:

```
%s [[ -2.48765602e-01 3.31598227e-01 6.30921033e-02 -2.81310530e-01
5.74140964e-03 1.62374420e-02 4.24863486e-02 1.03090398e-01
9.02270802e-02 -5.25098025e-02 3.58970400e-01 -4.59139498e-01
4.30462074e-02 -1.33405806e-01 8.06328039e-02 -5.95830975e-01
2.40709086e-02]
[-2.07601502e-01 3.72116750e-01 1.01249056e-01 -2.67817346e-01
5.57860920e-02 -7.53468452e-03 1.29497196e-02 5.62709623e-02
1.77864814e-01 -4.11400844e-02 -5.43427250e-01 5.18568789e-01
-5.84055850e-02 1.45497511e-01 3.34674281e-02 -2.92642398e-01
-1.45102446e-01]
[-1.76303592e-01 4.03724252e-01 8.29855709e-02 -1.61826771e-01
-5.56936353e-02 4.25579803e-02 2.76928937e-02 -5.86623552e-02
1.28560713e-01 -3.44879147e-02 6.09651110e-01 4.04318439e-01
-6.93988831e-02 -2.95896092e-02 -8.56967180e-02 4.44638207e-01
1.11431545e-02]
```

```

[-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
-3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
-3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
-8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
 3.85543001e-02]
[-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
-4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
-4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
-2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
-8.93515563e-02]
[-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
-4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
 5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
-8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
 5.61767721e-02]

[-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
 3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
-5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
 1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
-6.35360730e-02]
[-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
 2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
 4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
 1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
-8.23443779e-01]
[-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
 5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
-2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
-3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
 3.54559731e-01]
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
-1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
 1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
 3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
-2.81593679e-02]

```

```

[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
-2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
 9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
-1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
-3.92640266e-02]
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
 1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
 1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
 4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
 2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
 2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
 2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
-5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
 1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
-7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
-2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
 4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
-1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
-2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
 2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
-1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
 1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
 7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
 4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
 6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
 3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
-1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
-4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
 2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
 1.22106697e-01]]

```

Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features.

Bartlett's Test of Sphericity:

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

H0: All variables in the data are uncorrelated.

Ha: At least one pair of variables in the data are correlated.

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

Here, we can see that the p-value is small (0.0), so now we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

Performing KMO Test:

Measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

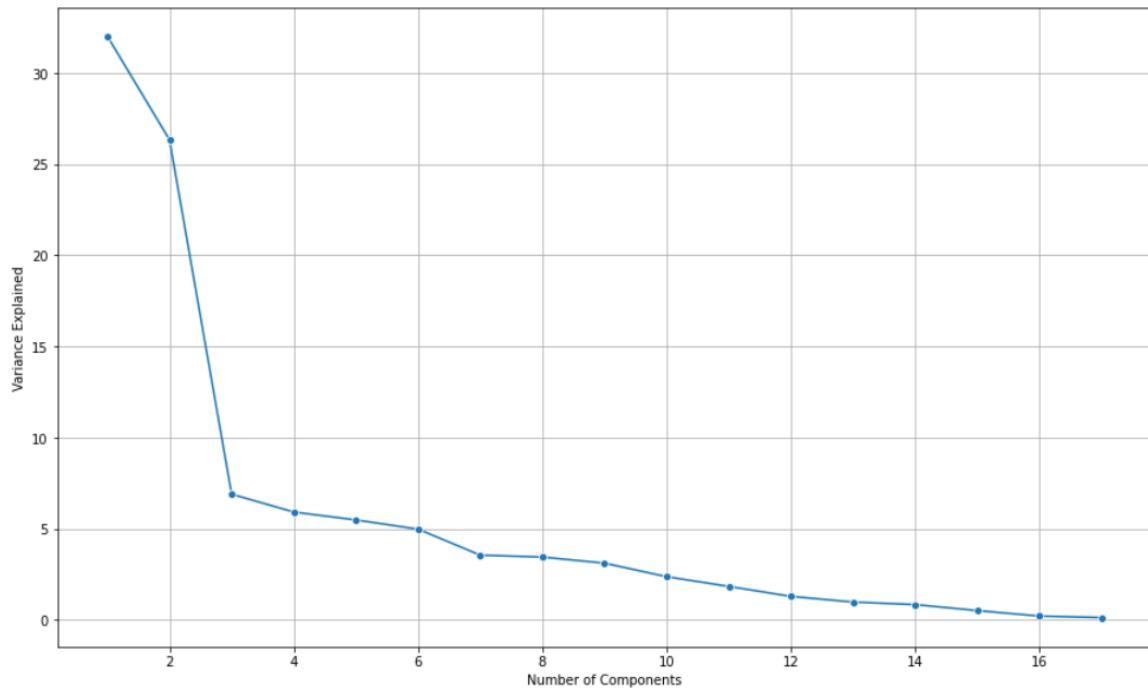
If MSA is less than 0.5, PCA is not recommended, since no reduction is expected.

But if the MSA > 0.7 then it is expected to provide a considerable reduction in the dimension and extraction of meaningful components.

0 . 8131251200373524

Here we can see that, MSA is 0.81 and that is greater than 0.7, so it is expected to provide a considerable reduction in the dimension and extraction of meaningful components.

Scree Plot: To get the number of components to be built.



We will take 6 PCA dimensions out of 17, because after that point, there is a continuous decline in the variance as displayed in the above scree plot.

Importing PCA from sklearn.decomposition and applying fit.transform to the scaled data:

```
array([[ -1.59285540e+00,   7.67333510e-01,  -1.01073452e-01,
       -9.21749413e-01,  -7.43975433e-01,  -2.98306010e-01],
       [-2.19240180e+00,  -5.78829984e-01,   2.27879802e+00,
        3.58891825e+00,   1.05999665e+00,  -1.77137392e-01],
       [-1.43096371e+00,  -1.09281889e+00,  -4.38092808e-01,
        6.77240527e-01,  -3.69613276e-01,  -9.60591686e-01],
       ...,
       [-7.32560596e-01,  -7.72352397e-02,  -4.05644759e-04,
        5.43162812e-02,  -5.16021117e-01,   4.68014245e-01],
       [ 7.91932735e+00,  -2.06832886e+00,   2.07356382e+00,
        8.52053973e-01,  -9.47754802e-01,  -2.06993727e+00],
       [-4.69508066e-01,   3.66660943e-01,  -1.32891512e+00,
        -1.08022562e-01,  -1.13217595e+00,   8.39893111e-01]])
```

PCA Components:

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
       0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
      -0.04252854,  0.31831287,  0.31705602,  -0.17695789,  0.20508237,
       0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425,  -0.08241182,  -0.04477866,
        0.41767377,  0.31508783,  -0.24964352,  -0.13780888,  0.05634184,
        0.21992922,  0.05831132,  0.04642945,  0.24666528,  -0.24659527,
      -0.13168986,  -0.16924053],
       [-0.0630921 ,  -0.10124906,  -0.08298556,  0.03505553,  -0.02414794,
      -0.06139299,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
        0.49972112,  -0.12702837,  -0.06603755,  -0.2898484 ,  -0.14698927,
        0.22674398,  -0.20806465],
       [ 0.28131053,  0.26781735,  0.16182677,  -0.05154725,  -0.10976654,
        0.10041234,  -0.15855849,  0.13129136,  0.18499599,  0.08708922,
      -0.23071057,  -0.53472483,  -0.51944302,  -0.16118949,  0.01731422,
        0.07927349,  0.26912907],
       [ 0.00574141,  0.05578609,  -0.05569364,  -0.39543434,  -0.42653359,
      -0.04345436,  0.30238541,  0.222532 ,  0.56091947,  -0.12728883,
        -0.22231102,  0.14016633,  0.20471973,  -0.07938825,  -0.21629741,
        0.07595812,  -0.10926791],
       [-0.01623744,  0.00753468,  -0.04255797,  -0.0526928 ,  0.03309159,
      -0.04345425,  -0.19119858,  -0.03000039,  0.16275545,  0.64105495,
        -0.331398 ,  0.09125552,  0.15492765,  0.48704588,  -0.04734001,
        -0.29811862,  0.21616331]])
```

Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

```
array([-0.25,  0.33,  0.06, -0.28,  0.01,  0.02,  0.04,  0.1 ,  0.09,
       -0.05,  0.36, -0.46,  0.04, -0.13,  0.08, -0.6 ,  0.02])
```

Above is the explicit form of first PC (in terms of eigenvectors) with 2 decimal places only.

Consider 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?

```
Cumulative values explained by eigen values : [ 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. ]
```

Eigen vectors are our principal components. Eigen values helps us to understand the quantum of variance which is being explained by our principal components.

