PGPDSBA Online FEB A 2021

Pavan Kumar R Naik

PGP-DSBA Online

Feb 2021

16/05/2021

Table of Contents

List of Figures

List of Tables

Problem 1 A:

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.

Introduction –

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of salaries of 40 individuals with respect to person's education qualification and occupation. This assignment will help in exploring the summary statistics, and hypothesis testing.

Data Description:

1. Education: Three categories (High School Graduate, Bachelor and Doctorate)
2. Occupation: Four categories (Adm-clerical, Sales, Professional or specialty and Executive or management)
3. Salary: Numerical Values (Range – 50103 to 260151)

Sample of Dataset:

	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

Table 1. Dataset Sample (Salary Data)

Summary of Dataset:

	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

Table 2. Dataset Summary (Salary Data)

Type of Variables:

```
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)
```

Table 3. Type of Variables (Salary Data)

Q1A 1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

Solution:

Formulation of hypothesis for conducting one-way ANOVA for education qualification w.r.t salary

- H0: Salary depend on education qualification
- Ha: Salary does not depend on education
- Confidence level = 0.05

Formulation of hypothesis for conducting one-way ANOVA for occupation w.r.t salary

- H0: Salary depend on occupation
- Ha: Salary does not depend on occupation
- Confidence level = 0.05

Q1A.2 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.

Solution:

To perform one-way ANOVA for education w.r.t the variable 'Salary', we apply the ANOVA formula in the Jupyter notebook and run the AOV table. We get following output:

	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

Table 4. Anova on Salary (Education)

From the above table, we find that the P value is less than 0.05, hence we reject the null hypothesis.

Q1A.3 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.

Solution:

To perform one-way ANOVA for occupation w.r.t the variable 'Salary', we apply the ANOVA formula in the Jupyter notebook and run the AOV table. We get following output:

	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

Table 5. Anova on Salary (Occupation)

From the above table, we find that the P value is greater than 0.05, hence we do not reject the null hypothesis.

Problem 1 B

Q1B.1 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.

Solution:

As seen from the below interaction plots, there seems to be moderate interaction between the two categorical variables.

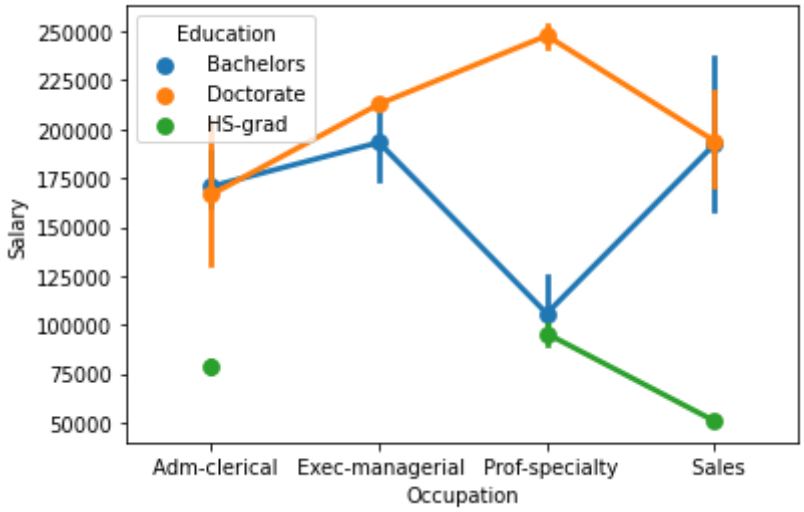

Fig 1. Point plot (Occupation vs Salary) with ci

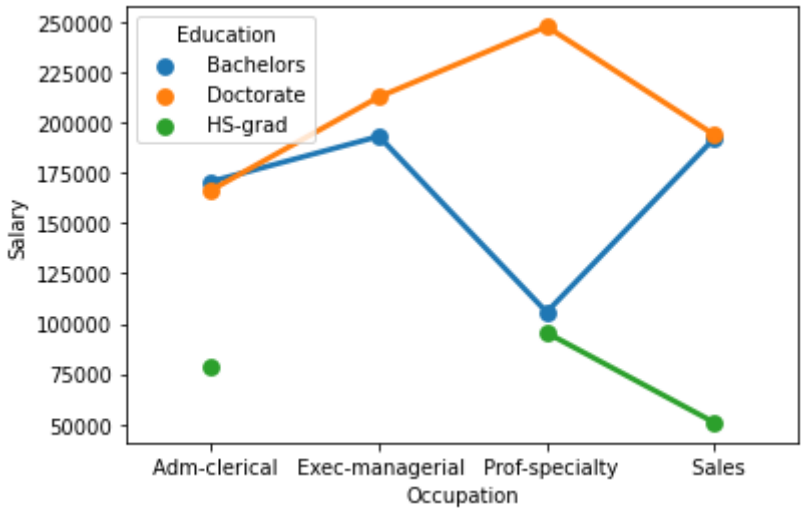

Fig 2. Point plot (Occupation vs Salary) without ci

As seen from the above interaction plots, there seem to be a moderate interaction between two categorical variables.

ADM-Clerical and Sales professional with Bachelors and Doctorate degree earn almost similar salary package.

Q1B.2 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?

Solution:

Formulation of hypothesis for conducting two-way ANOVA based on education and occupation w.r.t salary.

- H0: Salary depends on both categories - education and occupation
- Ha: Salary does not depend on at least one of the categories - education and occupation
- Confidence level = 0.05

	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

Table 6. Two-way Anova on Salary (Education and Occupation)

```
C(Education)                 True
C(Occupation)               False
C(Education):C(Occupation)   True
Residual                    False
Name: PR(>F), dtype: bool
```

Table 7. P Value

Considering both education and occupation, education is a significant factor as the P value is 0.05.

The interaction between education and occupation also influencing the Salary.

Q1B.3 Explain the business implications of performing ANOVA for this particular case study.

Solution:

We conclude that by doing ANOVA for the given dataset, Salary is depending on Occupation and there is a moderate interaction between Education and Occupation variables.

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.

Introduction:

The purpose of this whole exercise is to explore the dataset. Do the PCA. Explore the dataset using central tendency and other parameters. The data consists of names of various colleges. This particular case study is based on various parameters of various institutions. Expected to do the principal component analysis.

Data Description:

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

Sample of Dataset:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.
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	

Table 8. Dataset Sample (Education-post 12th)

Summary of 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	22.743887	12.391801	0.0	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

Table 9. Dataset Summary (Education-post 12th)

Type of Variables:

```
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)
```

Table 10. Dataset Variable type (Education-post 12th)

Q2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

Solution:

After importing all the relevant libraries on Jupyter notebook, we load the data set. We perform EDA to extract and see patterns in the given data set.

The given data set has a shape of (777, 18). Also, we check the top 5 rows of the data set then went on to see if there are any missing values in it – as per the output there are no missing 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

Table 11. Dataset Missing Values (Education-post 12th)

Then found that there are no duplicate values in the data frame.

Then did EDA using sweet viz to visualize the summary for each variable as well to underrated data

Univariate Analysis:

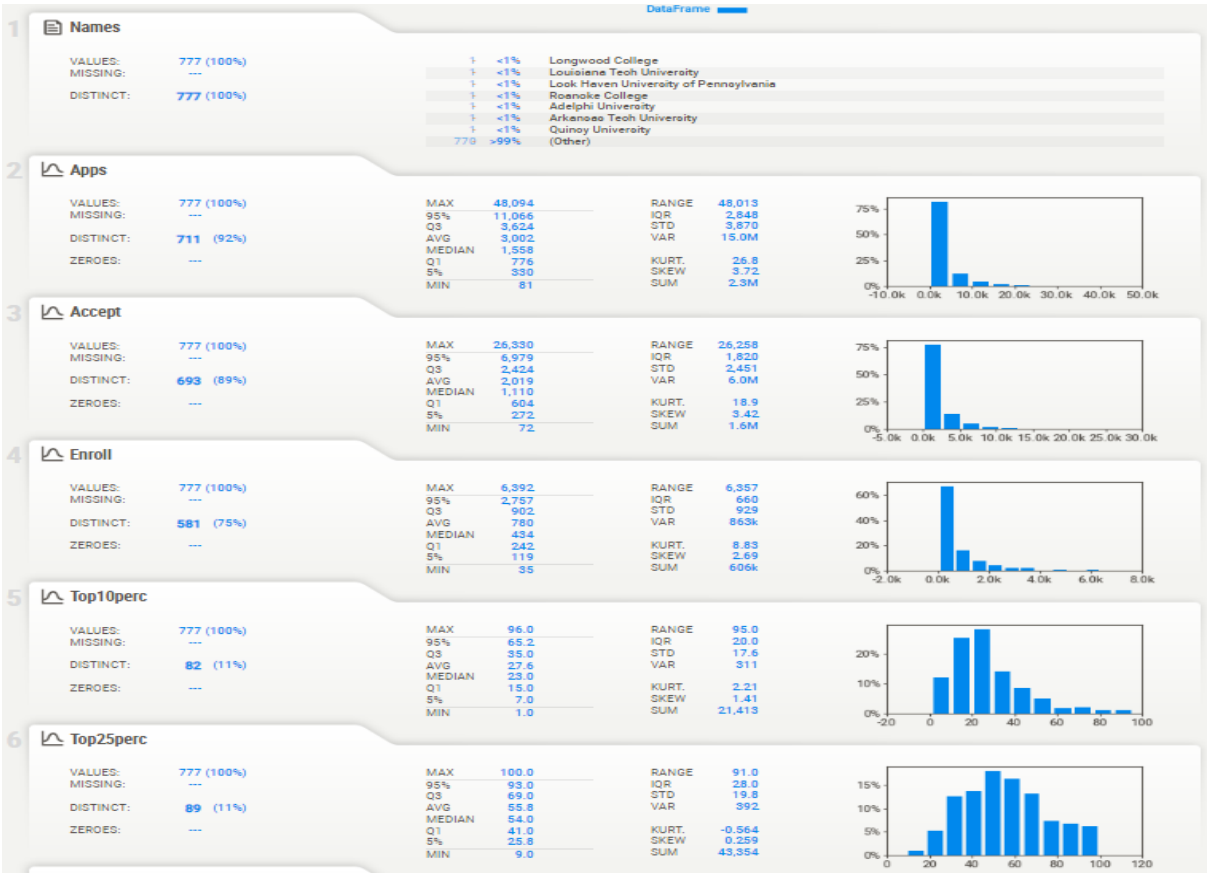

Fig 3.1. Sweet viz Univariate analysis

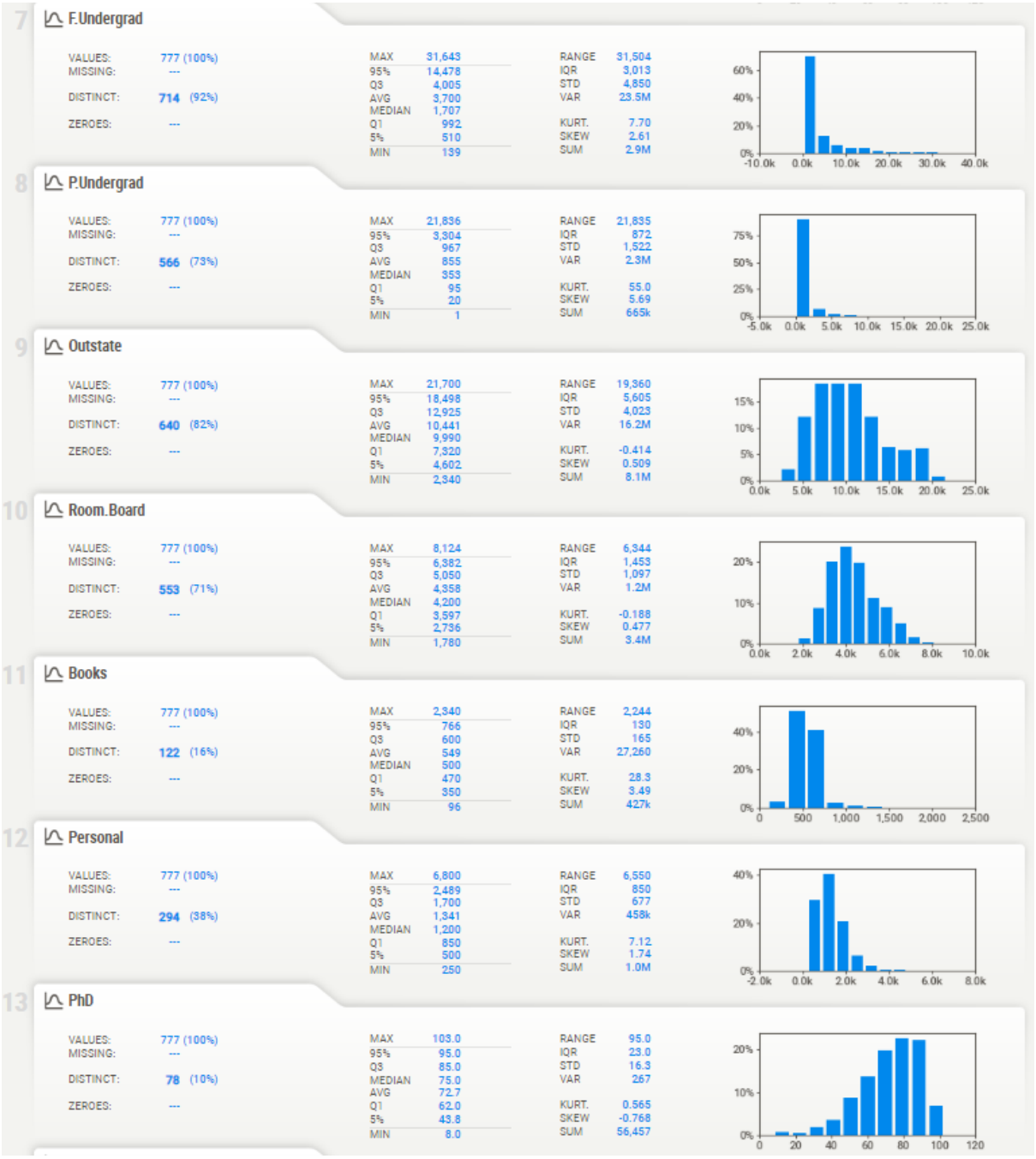

Fig 3.2. Sweet viz Univariate analysis

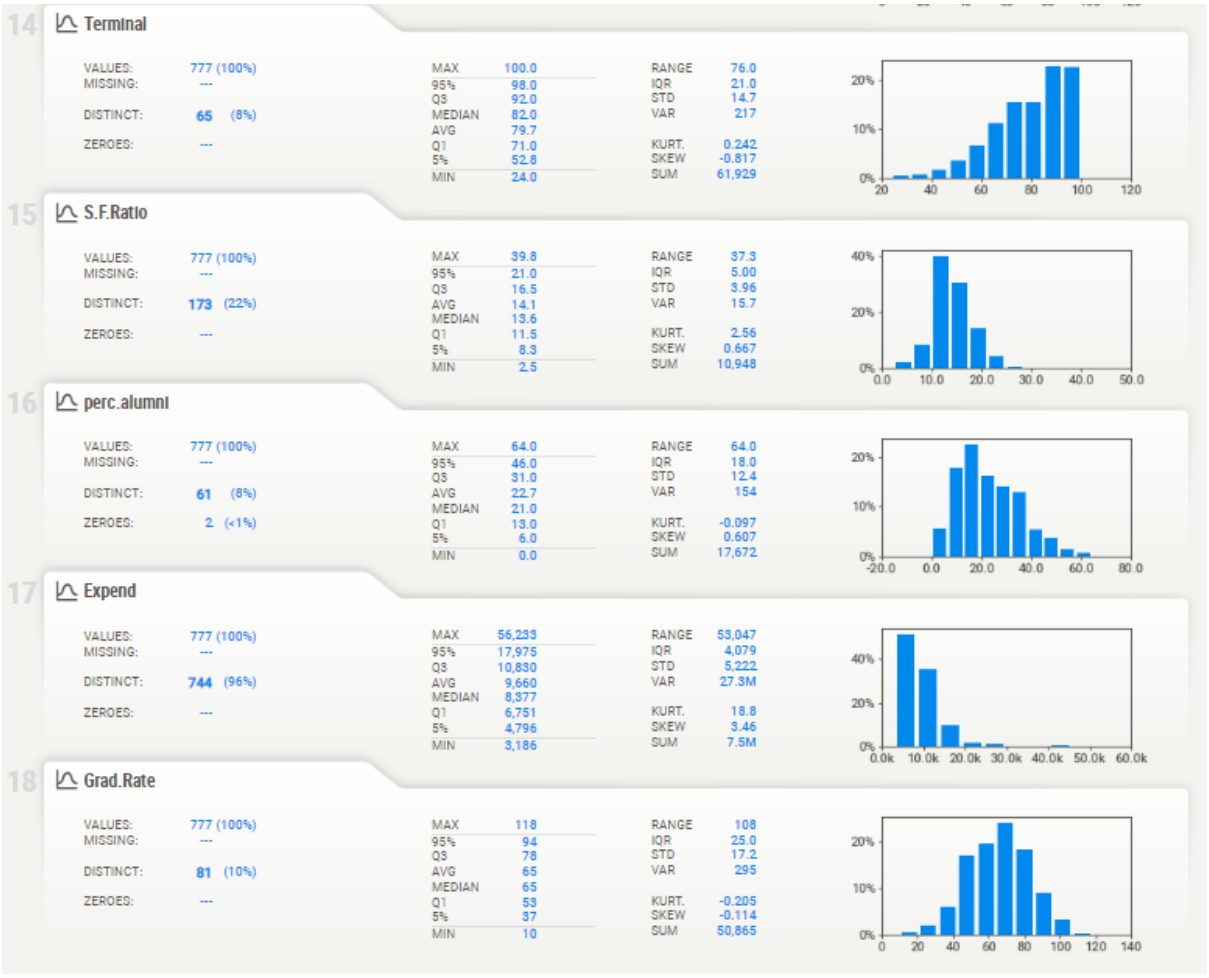

Fig 3.3. Sweet viz Univariate analysis

Insights:

- Variables Apps,Accept, Enroll, F.Undergrad, P.Undergrad, Top10perc, Books,Personal and Expend are highly right skewed
- Variables Top25perc, Outstate, Room board, Grad.Rate looks to be normally distributed
- Variables perc.Alumni, S.F ratio are near normal distribution
- Variables PhD, Terminal are left skewed
- Only variable Name is categorical variable

Multivariate Analysis:

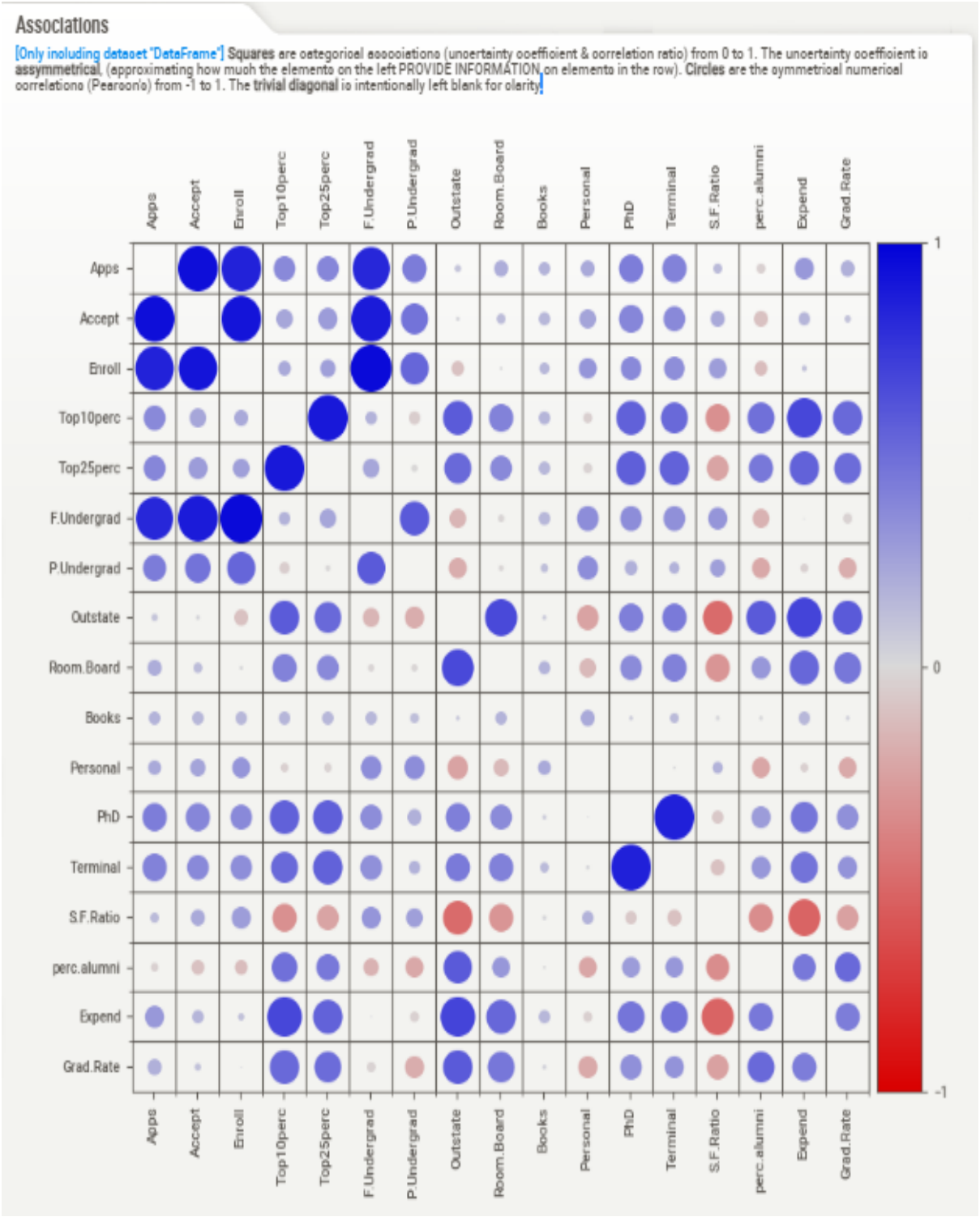

Fig 3.4. Sweet viz Multivariate analysis

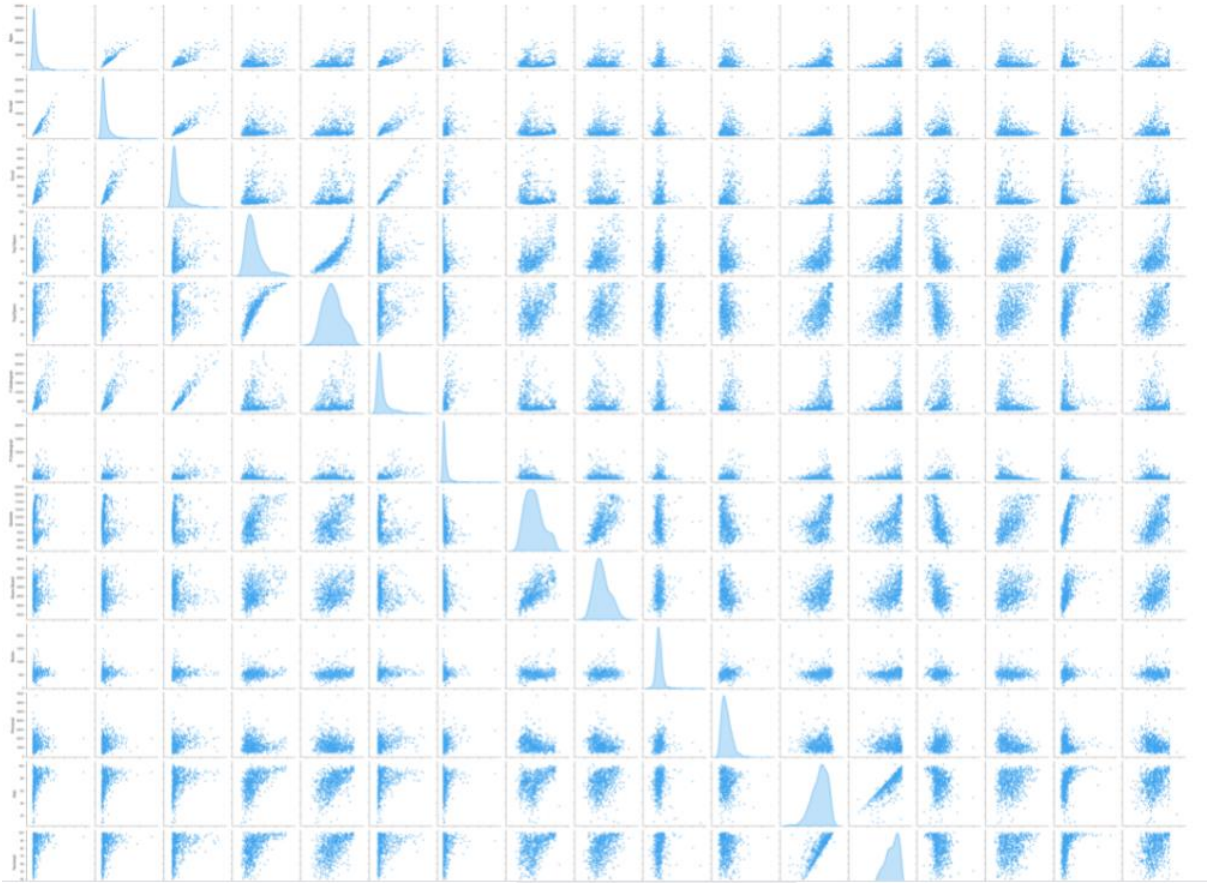

Fig 3.5. SNS Pairplot Multivariate analysis

Insights:

- Distribution of variables shows that most of the values are concentrated towards lowers side. Possibility of outliers
- F.undergrad, Apps, Accept and Enroll are highly correlated with each other
- PhD and Terminal are highly correlated with each other
- Top10Perc and Top25Perc are highly correlated with each other
- Outstate and expend are high correlated
- Outstate and S.F ratio are highly negatively correlated

Then we need to preprocess by removing the unwanted columns in this case its “Names”.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Ex
0	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	
1	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	1
2	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	
3	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	1
4	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	1

Table 12. Pre-processing (Education-post 12th)

Boxplot to identify the outliers.

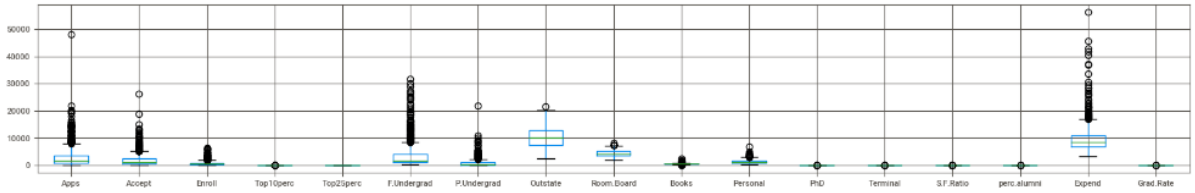

Fig 4. Boxplot

Q2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling.

Solution:

Yes, it is necessary to perform scaling for PCA. For instance, in given data set, applications and other variables are having values in thousands and few variables such as percentile is in just two digits. Hence the data in these variables are of different scales, it is tough to compare these variables.

The PCA calculates a new projection of the given data set and the new axis are based on the standard deviation of the variables. So, a variable with a high standard deviation in the data set will have a higher weight for the calculation of axis than a variable with a low standard deviation. By performing scaling, we can easily compare these variables.

We get the following output post we perform scaling using Z score.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R
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	1.013
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	-0.477
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	-0.300
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	-1.615
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	-0.553

Table 13. Z score (Education-post 12th)

Summary of data post scaling.

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

Table 14. Summary post Z score (Education-post 12th)

Inference we used is ZSCALER to standardize the data into single scale. Now all variables are in in between the scale of -2.5 to 12.5 with standard deviation of 1.0006.

Q2.3 Comment on the comparison between the covariance and the correlation matrices from this data. [on scaled data]

Solution:

Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. What sets them apart is the fact that correlation values are standardized whereas covariance values are not.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termin
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.3694
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.3375
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.3082
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.4911
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.5247
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.3000
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.1419
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.4079
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.3745
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.0999
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.0306
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.8495
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	1.0000
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	-0.1601
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.2671
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	0.4387
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	0.2895

Table 15. Correlation for scaled data (Education-post 12th)

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950	-0.030
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289	0.850
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682	1.001
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.160
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330	0.267
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319	0.439
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431	0.289

Table 16. Covariance for scaled data (Education-post 12th)

Q 2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

Solution:

Before scaling, let's plot a boxplot to check the outliers in all the variables. We get the following output:

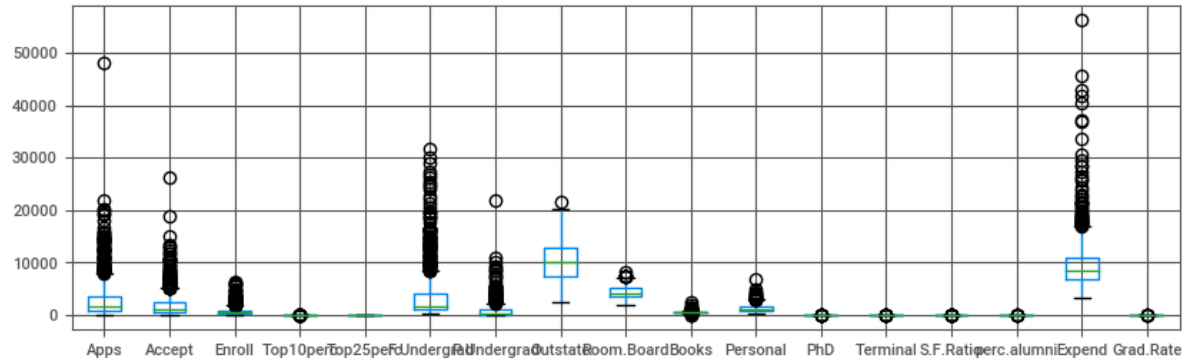

Fig 5. Boxplot before scaling

Post scaling, let's plot a boxplot to check the outliers in all the variables. We get the following output:

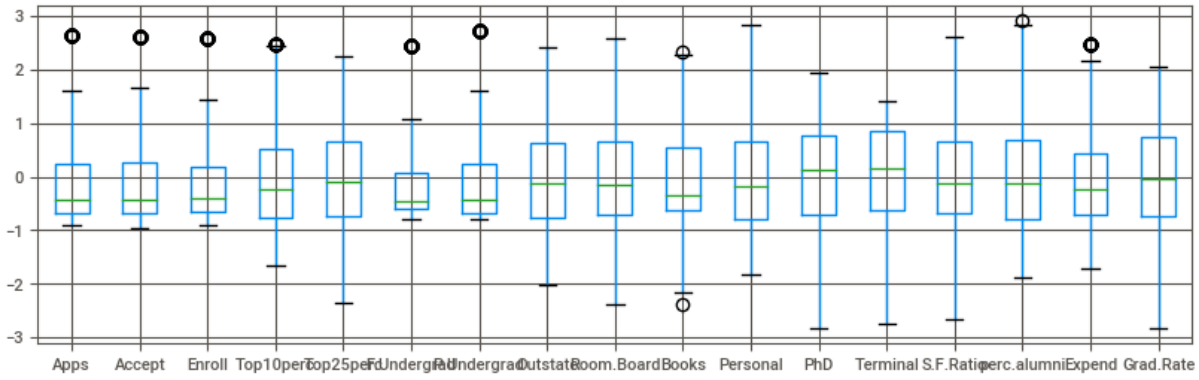

Fig 6. Boxplot after scaling

Insights:

- By scaling, all variables have the same standard deviation, thus all variables have the same weight and thus resulting in PCA calculating relevant axis
- Before scaling, we only had one variable with no outliers (top25 perc); Post scaling, we have multiple variables with negligible outliers – this is achieved by normalizing the scale of the variables

Q2.5 Extract the eigenvalues and eigenvectors. [print both]

Solution:

We can extract the above represented eigenvalues and eigenvectors using covariance matrix.

The below snapshot represents extracted eigenvalues and eigenvectors:

```
 Eigen Values
%s [0.03891348 0.05597992 0.07466871 0.12376406 0.13603844 0.14684496
 0.24563729 0.38222641 0.44319291 0.55450358 0.58709565 0.76549205
 0.8977433  0.9966849  1.10030644 4.82973672 5.64307841]
```

Fig 7. Eigen Values

```
Eigen Vectors
%s [[-1.51051724e-01  5.73869368e-01  2.54721171e-02  3.50002377e-01
   4.76265776e-01 -2.73993248e-02 -6.79408155e-02 -1.34049566e-01
   1.84655032e-01 -3.41030317e-02  1.23782113e-02  4.76414519e-02
  -2.28743180e-01  1.02559773e-01  9.77100175e-02 -3.24930495e-01
  -2.42671239e-01]
 [ 4.52766958e-01 -6.43625404e-01 -4.08143058e-02  1.12837998e-01
   2.08677137e-01 -1.27528369e-01 -2.86891699e-02 -1.23207526e-01
   1.89697047e-01 -1.02521665e-01 -1.41529768e-03  3.31338141e-02
  -2.02792107e-01  1.21914245e-01  1.25144023e-01 -3.57755851e-01
  -2.08095876e-01]
 [-7.50067816e-01 -2.58381892e-01  3.37484396e-02 -2.25003975e-01
  -2.65981931e-01 -1.80558174e-02 -2.29745788e-02 -4.79563882e-02
   5.20184210e-02 -1.34762063e-01  7.92830517e-03 -3.89761143e-02
  -1.72168365e-01  1.42497171e-02  9.44419384e-02 -3.95824297e-01
  -1.64564266e-01]
 [ 5.89947774e-02 -5.31897461e-02  7.23553559e-01 -3.22924466e-02
   1.62488072e-02  4.57358763e-02 -6.57319491e-03 -7.14429611e-02
  -1.10851590e-01  2.89094711e-01  2.58267694e-01 -8.37673857e-02
  -1.45905144e-01 -3.75563233e-01 -7.23866450e-02  7.53900839e-02
  -3.44633526e-01]
 [-1.47356588e-02 -3.70257583e-03 -6.58266244e-01  2.64582929e-02
  -3.47432747e-02 -1.58456329e-01 -1.32078205e-01 -4.53255044e-02
  -1.89924670e-01  3.36249057e-01  2.34717438e-01 -2.14918233e-02
  -1.20536687e-01 -4.27876370e-01 -4.63368319e-02  3.67211412e-02
  -3.37858398e-01]
 [ 4.51829780e-01  4.13880625e-01 -1.05340454e-02 -3.50240396e-01
  -5.20754661e-01  7.88826178e-02 -3.63762678e-02  1.12660606e-02
  -1.41252801e-04 -1.22385171e-01  2.79162755e-02 -5.49956869e-02
  -1.15073146e-01  1.46165800e-02  8.72397333e-02 -4.06243667e-01
  -1.34287678e-01]
 [ 4.97285352e-03 -3.24986907e-02  3.82640339e-02  1.01785907e-01
   1.61437628e-01 -3.58599650e-02  1.89391557e-01  4.23776360e-01
  -7.36103386e-01  5.41905661e-02  9.36586774e-02 -5.16448338e-02
   1.32038801e-01  2.07265372e-01  3.86964803e-02 -3.54916637e-01
  -1.45128920e-02]
 [-4.74030188e-03  9.51907262e-02  2.55089170e-03 -2.23348292e-01
  -7.53206365e-03 -5.57302446e-01  6.09931131e-01 -1.87206448e-01
  -1.46112057e-02  2.38893511e-02 -1.04399025e-01 -1.39668813e-02
  -4.29684243e-02  2.53851713e-01  2.05908405e-02  2.37362415e-01
  -2.97304568e-01]
```

Fig 8. Eigen Vectors

Q2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features

Solution:

For performing PCA, we need to follow below steps:

- Step 1: Generate the covariance matrix
- Step 2: Get eigenvalues and eigenvector
- Step 3: View Scree Plot to identify the number of components to be built
- Step 4: We can perform PCA on the scaled data set by importing PCA from sklearn.decomposition

We get following component output:

```
array([[-1.73690053, -1.59813534, -1.5427997 , ..., -0.67443639,
         6.6604872 , -0.62212047],
       [ 0.78652267, -0.33203883, -1.37926794, ..., -0.14322329,
        -1.08947808,  0.63056856],
       [ 0.0913572 ,  2.12932938, -0.60242058, ...,  0.37332843,
         1.41416131, -1.31545432]])
```

Fig 9. PCA on scaled data

We do load of each feature on the components.

```
array([[ 0.24267123,  0.20809587,  0.16456427,  0.34463353,  0.3378584 ,
         0.13428769,  0.01451289,  0.29730457,  0.2511921 ,  0.09356817,
        -0.04846688,  0.32466754,  0.32050994, -0.17847668,  0.19861754,
         0.34015699,  0.24864478],
       [ 0.32493047,  0.35775585,  0.39582431, -0.07539009, -0.03672113,
         0.40624369,  0.35491663, -0.23736241, -0.12378904,  0.10601539,
         0.23546922,  0.07065173,  0.05966643,  0.24783489, -0.24326186,
        -0.13574787, -0.16060776],
       [-0.09774393, -0.1251541 , -0.09442148,  0.0723851 ,  0.0463486 ,
        -0.08720429, -0.03870995, -0.02057768,  0.02607806,  0.71355757,
         0.52183311, -0.05730359, -0.03741349, -0.25838406, -0.10991406,
         0.17290996, -0.23102783]])
```

Fig 10. PCA Components

PCA scores in Data frame

	principal component 1	principal component 2	principal component 3
0	-1.736901	0.786523	0.091357
1	-1.598135	-0.332039	2.129329
2	-1.542800	-1.379268	-0.602421
3	3.181988	-2.993983	0.335459
4	-1.785882	-0.202227	2.730809

Table 17. PCA Scores

Principal component Data frame

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ra
0	0.242671	0.208096	0.164564	0.344634	0.337858	0.134288	0.014513	0.297305	0.251192	0.093568	-0.048467	0.324668	0.320510	-0.178
1	0.324930	0.357756	0.395824	-0.075390	-0.036721	0.406244	0.354917	-0.237362	-0.123789	0.106015	0.235469	0.070652	0.059666	0.247
2	-0.097744	-0.125154	-0.094421	0.072385	0.046349	-0.087204	-0.038710	-0.020578	0.026078	0.713558	0.521833	-0.057304	-0.037413	-0.258

Table 18. PC DF

Heatmap

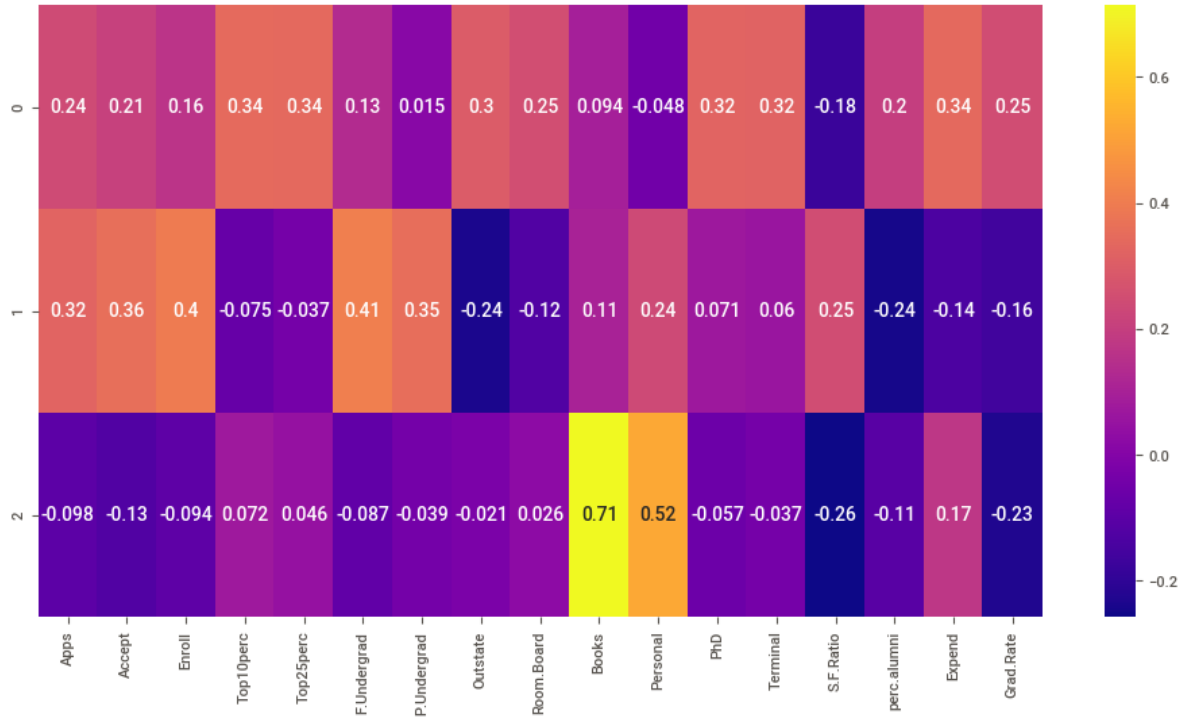

Fig 11. Heatmap

Communality (Items correlates with all other items)

```
Apps            0.665346
Accept          0.691006
Enroll          0.654810
Top10perc       0.492409
Top25perc       0.420928
F.Undergrad     0.627736
P.Undergrad     0.408139
Outstate        0.555245
Room.Board      0.401059
Books           0.913141
Personal        0.805769
PhD             0.452623
Terminal        0.417590
S.F.Ratio       0.684696
perc.alumni     0.551793
Expend          0.648815
Grad.Rate       0.640280
dtype: float64
```

Fig 12. Communality

Q2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

Solution:

If we sort the eigenvectors in descending order with respect to their eigenvalues, we will have that the first eigenvector accounts for the largest spread among data, the second one for the second largest spread and so on.

```
array([[ 1.74,  1.6 ,  1.54, -3.18,  1.79,  0.55, -0.23, -1.9 , -0.8 ,
         2.84, -1.93, -2.2 ,  0.09, -0.88,  2.2 ,  1.51, -5.23,  2.22,
         2.03,  2.98, -0.17, -0.42,  1.8 , -1.14, -0.69,  3.46, -1.27,
        -1.35,  1.65, -1.03,  0.94, -1.16,  2.66,  1.97,  0.03,  1.15,
        -3.82, -3.74,  0.47, -1.36, -0.45,  0.06,  1.71,  1.34, -0.94,
         3.65,  2.2 ,  0.4 , -1.27,  0.34,  2.04,  1.47,  3.22,  3.46,
        -1.15,  1.17,  1.05,  3.14,  1.28, -4.92, -5.31, -0.72,  0.56,
        -0.82, -4.21,  1.61,  3.93,  2.29,  0.52, -1.33, -6.6 , -4.94,
        -4.27,  0.57, -1.28,  1.49,  0.64,  0.43, -1.15,  0.1 , -0.09,
         1.53,  2.6 , -0.06, -0.21,  3.38, -4.19, -5.29,  0.62,  1.72,
         0.73, -4.17,  1.31,  1.3 , -1.66,  1.63, -0.42,  1.74,  1.73,
        -0.07,  1.86, -0.28,  2.33,  2.27,  1.41,  3.29, -2.7 , -0.08,
        -1.54,  0.08, -0.2 ,  1.01,  2.32,  2.78, -4.78, -2.15,  1.41,
        -2.35, -2.18,  3.19, -0.29,  0.83, -3.92, -4.31,  0.21,  1.05,
         1.85, -0.29,  1.04,  0.12, -0.18,  0.37,  0.32,  0.65,  1.86,
         2.63,  0.6 , -4.51, -3.44, -2.17, -3.47, -1.68,  2.58,  0.64,
        -5.7 ,  2.27,  2.66,  1.27,  1.69, -3.97, -0.14, -0.6 , -2.44,
         2.39,  3.01,  1.45,  3.03,  2.26, -6.53, -4.6 ,  2.74,  3.43,
        -2.42, -2.36, -2.85,  3.91,  1.39,  1.61,  3.81,  1.87,  1.93,
        -1.47, -3.67, -0.25, -6.64, -2.21, -0.71,  1.53,  3.79,  1.93,
         1.69,  1.52,  1.32,  2.64, -1.28, -0.51,  0.08,  0.64,  1.34,
         2.77, -0.53, -6.32,  1.47, -0.39,  0.59,  1.23, -2.64,  3.51,
         3.04,  3.2 , -1.07, -1.03,  1.74, -2.28,  3.33, -2.13,  1.61,
         2.81,  1.6 ,  0.23,  2.4 ,  2.04,  0.71, -2.64,  1.53,  2.86,
         1.71,  1.94, -0.32, -4.11,  0.91, -6.06, -3.81,  1.72,  1.85,
        -3.28,  4.67, -0.79,  0.42,  1.04, -2.4 ,  2.24,  3.32,  1.37,
         1.87,  2.14,  1.99, -3.8 ,  0.31,  0.16, -1.55,  1.24, -3.85,
        -1.13, -1.29,  0.68,  0.02,  2.36,  0.28, -0.97, -6.45, -5.37,
         2.21, -0.81, -0.87, -1.37, -2.76, -2.07, -1.45, -1.39, -0.8 ,
         0.03,  1.97,  1.37,  4.85,  4.15, -0.02,  0.94, -1.5 ,  0.1 ,
        -2.02,  1.39,  0.85,  1.03, -1.99,  3.13,  0.1 , -1.53, -1.77,
        -2.32,  3.21,  2.78,  1.58, -1.33, -6.11,  2.27,  2.88, -1.77,
        -0.14,  2.57,  2.05,  1.72, -3.33,  1.11,  0.65,  0.97, -2.1 ,
         2.2 , -0.66, -3.51,  1.27, -2.02,  2.47,  3.3 ,  3.13,  2.93,
        -2.59, -1.02, -0.61, -4.7 ,  0.28,  1.4 ,  1.85, -2.04,  2.99,
         1.94,  2.21,  1.85, -1.4 ,  0.88,  2.57,  1.46,  1.5 ,  1.5 ,
        -0.18,  2.36, -1.58, -1.93, -0.8 , -2.01, -0.95,  0.5 ,  1.46,
         3.08, -3.  ,  2.18,  3.28,  1.04, -1.49, -1.5 ,  2.53,  2.87,
        -1.08,  0.02, -2.31,  0.32, -1.1 , -0.54, -0.12,  0.67,  0.98,
         0.01,  2.19,  0.66, -5.83,  3.79,  2.15,  2.8 ,  2.38, -1.48,
         2.04,  0.73,  0.36,  4.21, -0.2 , -1.89, -2.55, -0.73,  3.64,
         0.34,  1.6 ,  0.22, -0.63,  1.81,  1.38,  0.39,  3.52,  2.97,
         3.69, -0.7 ,  0.68,  0.75,  1.38,  0.68,  3.74,  1.93, -0.73,
         0.01,  2.1 ,  3.68, -3.98,  3.2 ,  2.64,  2.44,  3.51, -0.24,
         1.24, -0.47, -0.08,  3.12, -1.57,  1.25,  0.1 ,  1.77, -0.55,
        -1.3 ,  0.12, -6.12,  2.12, -1.62,  2.81,  2.14, -3.64,  1.85,
        -0.59,  1.38,  0.57,  0.08, -1.96,  0.42,  0.07,  3.13,  2.19,
         1.6 , -6.71,  0.76,  3.24,  0.5 , -4.34, -3.23, -1.59, -1.12,
        -1.2 , -2.37,  1.98,  2.99, -0.43,  0.26,  1.69,  2.15, -2.45,
        -0.83,  1.12, -0.22,  2.57, -3.02, -4.23,  3.17,  2.96,  1.08,
         1.87,  1.6 ,  2.86, -2.72,  1.31,  1.98, -2.1 ,  2.95, -0.6 ,
        -6.34, -0.95, -2.03,  0.14,  0.78, -0.64,  0.96,  1.99, -0.2 ,
        -2.06, -3.18, -0.62, -4.02, -3.21, -0.78, -1.07,  2.1 , -0.62,
         0.93,  1.99,  2.07, -1.57, -0.11,  0.87, -3.17, -0.49, -0.88,
         0.37,  2.16,  0.08,  1.97,  2.92,  1.58,  0.03,  1.75,  1.99,
        -1.24, -0.4 , -2.87, -1.14,  2.05,  1.45, -1.04, -1.93,  0.59,
         1.2 , -0.09,  1.78, -0.11,  1.23,  0.25, -1.64, -2.68, -3.72,
```

```
 0.86,  1.78, -4.19, -1.04, -1.04, -0.73,  0.45,  0.65,  1.79,
-1.08,  3.23, -1.56,  0.47,  2.54, -2.69, -4.44,  2.22,  2.13,
 3.57,  2.55,  1.96, -1.84,  3.44,  1.1 ,  2.74,  3.52,  2.18,
-1.73,  2.06, -0.88,  3.23, -0.16, -0.48,  0.12, -2.92,  3.05,
-1.75, -0.77,  0.44, -0.27,  4.05,  1.97,  1.75, -1.18, -2.62,
 0.76, -1.33, -2.05, -3.49, -3.59, -2.72,  1.11, -0.17,  1.65,
 0.71,  0.55, -1.92, -0.11,  0.72,  1.15,  1.17, -1.33, -2.91,
-3.76,  2.51,  2.24, -0.14,  3.37, -2.56,  1.44, -0.65,  2.31,
 0.86,  1.84,  1.03,  1.06,  3.01, -1.63, -1.81,  2.2 , -4.29,
-1.48,  0.71, -3.54, -4.5 ,  2.91,  0.89,  2.21, -4.17,  1.55,
-0.76,  0.36, -5.27, -5.25, -0.47,  2.49, -5.8 , -2.4 , -2.94,
-1.17, -1.64, -2.13, -1.01,  0.8 ,  1.38, -0.25, -4.07, -2.6 ,
-0.01, -0.91, -3.15, -1.87,  1.4 , -1.82,  1.12,  0.07,  1.83,
 3.32,  3.17,  0.33, -2.05, -1.52,  0.89, -4.28, -5.59,  0.71,
 0.06, -2.93,  0.59, -1.63, -0.92,  1.67,  2.91,  2.24, -1.61,
 0.46, -2.15,  0.95, -5.  , -0.13,  0.19,  0.32,  0.15,  0.2 ,
 0.82,  1.52,  0.13, -5.27, -1.12, -1.42, -7.14, -3.23, -0.7 ,
-1.92, -1.63, -2.49, -5.67, -2.07, -0.69,  3.8 , -1.95,  3.6 ,
-1.48, -0.86, -5.06,  2.8 ,  3.51,  0.97, -0.21,  0.39, -1.3 ,
 2.52, -2.5 ,  0.35,  2.12, -2.08, -3.14, -1.2 , -0.99, -2.52,
-5.13, -4.1 ,  1.24,  2.21,  2.23,  1.1 ,  1.98, -3.56,  0.51,
 0.54,  2.42, -2.06,  2.08,  4.8 , -1.58, -5.9 , -4.63, -3.51,
-0.59,  2.94, -2.25,  2.61,  0.78,  1.88,  4.78, -1.92,  0.36,
-4.65,  1.13,  1.41, -0.08, -1.7 , -3.83, -1.14, -2.19, -5.98,
 3.65,  2.66,  3.49,  2.28, -5.06, -2.83,  1.04,  2.45, -4.58,
 0.7 ,  4.24,  0.36,  2.35, -0.89, -0.72,  1.19,  2.84, -0.51,
 2.21,  1.45,  0.17,  0.83, -0.79, -1.97, -2.73,  1.05, -2.56,
-1.28, -0.51, -0.7 ,  0.23, -2.06,  0.8 ,  2.58, -5.28,  0.23,
 1.66,  2.88,  0.85,  2.6 , -1.7 , -1.68, -2.91,  3.58, -0.27,
 0.67, -6.66,  0.62]])
```

Fig 13. Explicit form of first PC

Q2.8 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?

Solution:

Cumulative values is the sum of all eigen values. To find the variance explained each component we should divide each component's eigenvalue by the cumulative values.

By knowing the cumulative values, we will know the percent of information captures in each principal component and decide the optimum numbers.

From the below screenshot of cumulative values of the eigenvalues, we can see that around 8 principal components explained over 90% of the variance. Thus, the optimum number of principal components can be 8.

```
Cumulative Variance  [ 33.15185743  61.52550945  67.98957029  73.84487717  79.11892366
  83.61602283  87.06508197  90.32266981  92.9263317   95.17182864
  96.61489423  97.47757639  98.27677262  99.00385952  99.44252192
  99.77139178 100.        ]
```

Fig 14. Cumulative Value

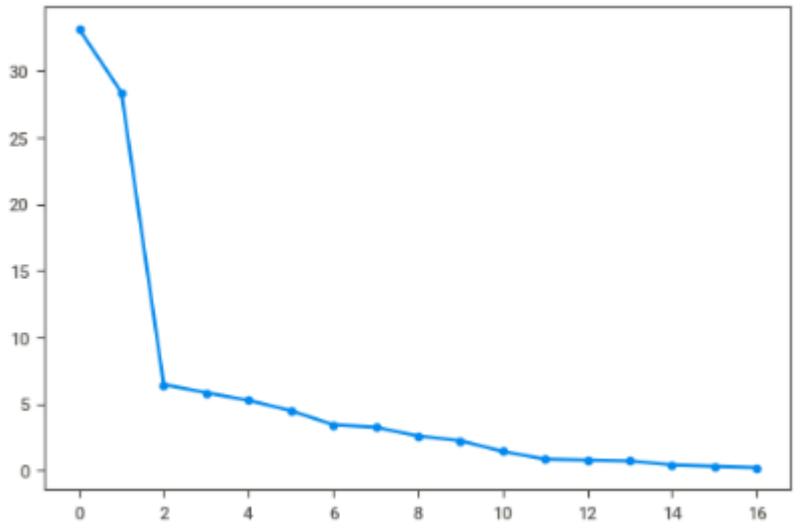

Fig 15. Plot for Scree test

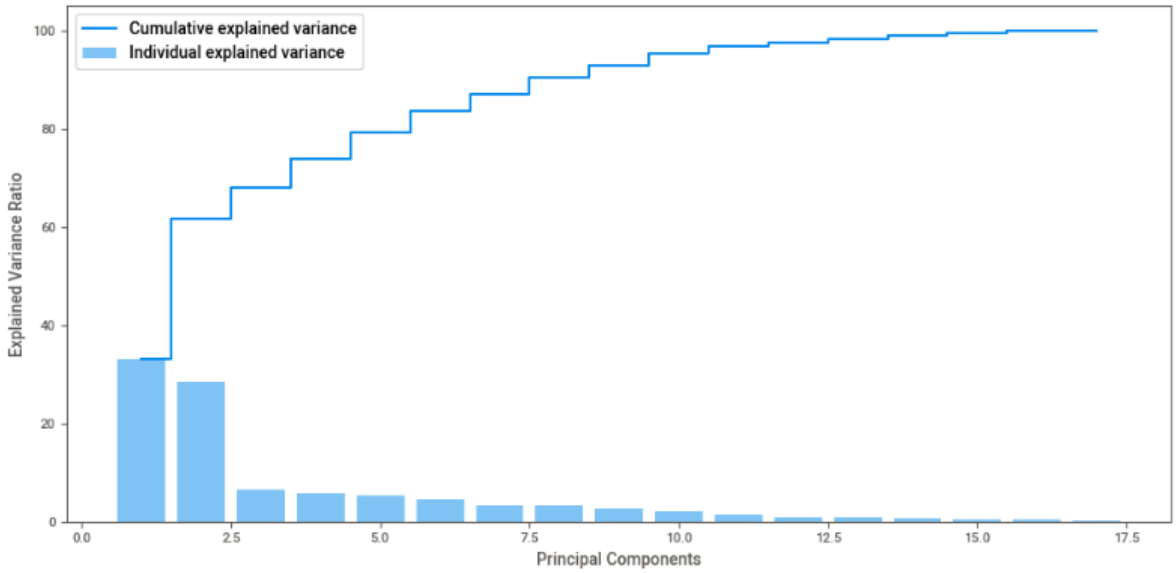

Fig 16. PC Vs Variance ratio

From the above cumulative graph, we can find that first two PCA components is picking up around 60%.

Furthermore, eigenvectors indicate coefficient of the features or numerical columns, the direction of the principal components, we can multiply the original data by the eigenvectors to re-orient our data onto the new axes.

Q2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?

Solution:

The three Principal components (PC0, PC1 and PC2) created are free from multicollinearity

Just three PCA (out of 17) components is picking up around 68 % of variability.

PC0 explains most of variables at average level of .22 with good explanate for top 10 perc, top 20 perc , expend, PhD, terminal, outstate variables.

PC1 has good explanation for f.undergrad ,enroll ,accept, p undergrad , accept and apps

PC2 has highest explanation for Books and personal.

The highest communality variable is Personal with 81% communality

Thus, as far as business implication of using PCA is concerned, in this case, we are reducing a high dimensional space (with 17 variables) and converting it to a lower dimensional space without (theoretically) losing much of the explanatory power.

Thank you,

Pavan Kumar R Naik