Advanced Statistics ASSIGMENT

Submitted By

Vinit Sharma

Table of Contents

Figures of Contents

Problem 1A: Salary Analysis

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.]

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

Hypothesis on One Way ANOVA for Education:

- **Null Hypothesis:** Mean salary of each level of education is equal.
- **Alternate Hypothesis:** At least one of the means of salary for level of education is not equal.

Hypothesis on One Way ANOVA for Occupation:

- **Null Hypothesis:** Mean salary of each level of occupation is equal.
- **Alternate Hypothesis:** At least one of the means of salary for level of occupation is not equal.

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

One-way ANOVA on Salary with respect to Education:

- **Null Hypothesis:** Mean salary of each level of education is equal.
- **Alternate Hypothesis:** At least one of the means of salary for level of education is not equal.

One way ANOVA result:

	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

FIGURE 1 ONE-WAY ANOVA ON SALARY WITH RESPECT TO EDUCATION

As P-value is less than alpha (0.05), we have enough evidence to **reject** the null hypothesis.

1.1.3 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.

One-way ANOVA on Salary with respect to Occupation:

- **Null Hypothesis:** Mean salary of each level of education is equal.
- **Alternate Hypothesis:** At least one of the means of salary for level of education is not equal.

One way ANOVA result:

	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

FIGURE 2 ONE-WAY ANOVA ON SALARY WITH RESPECT TO OCCUPATION

As P-value is greater than alpha (0.05), we have **not** enough evidence to reject the null hypothesis.

1.1.4 If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded)

Null Hypothesis is rejected for the Education related ANOVA test (2). ANOVA explains about the test result is significant or not. It does not provide the clear picture of" where" on the statistical significance. To measure the interpretability of statistical significance of our ANOVA test, we perform Tukey Test. It is also used to find out which specific groups' means are different.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7541.1439	79006.9894	True
Bachelors	HS-grad	-90114.1556	0.001	-132035.1958	-48193.1153	True
Doctorate	HS-grad	-133388.2222	0.001	-174815.0876	-91961.3569	True

FIGURE 3 TUKEY HSD ANALYSIS FOR EDUCATION

Based on the result, mean count of the Salary differs for all levels of education. For the pairs of bachelors/doctorate means diff value is higher than bachelors/ HS-grad, Doctorate/HS-grad. All group pairs results are **rejected**.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

FIGURE 4 TUKEY HSD FOR OCCUPATION

Based on the result, mean count of the Salary for all levels of occupation are differs. All group pairs results are not **rejected**.

Problem 1B: Salary Analysis

1.2.1 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.[hint: use the 'pointplot' function from the 'seaborn' function]

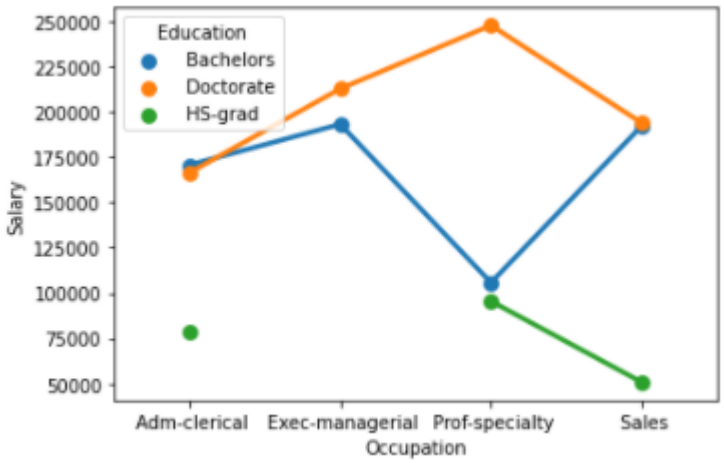

FIGURE 5 INTERACTION PLOT BETWEEN EDUCATION AND OCCUPATION

Some points related to above plot has shown below.

1. Bachelors and Doctorate level of education has earned good in adm-clerical occupation.
2. HS-grad education level people have not shown any result for exec-managerial.
3. For an exec-managerial occupation, Doctorate has earned a bit higher compared to bachelors.
4. Prof-speciality with Doctorate earned the most in comparison with Bachelors and HS-grad. Looks like for the occupation of prof-speciality with bachelors and HS-grad does not earn much.
5. HS-grad in sales occupation has lowest salary earned compared to other education level people for same roles. Relatively, Doctorate and Bachelors have same kind of salary portion.
6. HS-grad people salary range is lowest compare to other education level people in all four-occupation level.

1.2.2 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?

Two-way ANOVA on Salary with respect to Occupation and Education:

- Null Hypothesis: Mean salary of each level of education and Occupation is equal.

- **Alternate Hypothesis:** At least one of the means of salary for level of education and Occupation is not equal.

Two-way ANOVA result (No Interactions):

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	2.284576	9.648715e-02
C(Education)	2.0	9.695663e+10	4.847831e+10	29.510933	3.708479e-08
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

FIGURE 6 TWO-WAY ANOVA RESULT (WITHOUT INTERACTIONS)

Two-way ANOVA result (With Interactions):

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	5.277862	4.993238e-03
C(Education)	2.0	9.695663e+10	4.847831e+10	68.176603	1.090908e-11
C(Education):C(Occupation)	6.0	3.563950e+10	5.939916e+09	8.353494	2.643089e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

FIGURE 7 TWO-WAY ANOVA RESULT (WITH INTERACTIONS)

Based on result which includes effect of interactions, we can notice that P value has reduced for both occupation and education compare to non-interactive result.

As P-value is less than alpha (0.05), we have enough evidence to **reject** the null hypothesis.

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

Business implications of performing ANOVA for this particular case study is to understand how salary of a person varied based on education and occupation level. Specific occupation such as Prof-speciality with Doctorate earned the most in comparison with Bachelors and HS-grad. This case study also reflects about interaction effects between education and occupation. Bachelors and Doctorate level of education has earned good in adm-clerical occupation. HS-grad people salary range is lowest compare to other education level people in all four-occupation level. For a certain profession, there is not much value of higher qualification as salary range is same for example sales occupation.

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.

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

To initialize our problem statement, First, we need to understand the intent of the problem, its variables and datatype. We need to perform some operation such as find the duplicate values, unique value, missing value on the dataset. Outlier identification and five-point summary on the dataset are also an essential step on EDA.

EDA

Univariate Analysis

Entire analysis is based on the single variable prospective. The statistical description of the numeric variable, histogram or distplot to view the distribution andthe box plot to view 5-point summary and outliers if any.

	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

FIGURE 8 FIVE POINT SUMMARY ANALYSIS

Data has total 770 count and 18 variables. Based on initial analysis, dataset has no duplicate and null values. We have one object datatype variable “Names”. Each variable has min and max value with a very wide range.

	min	max	IQR Range	Upper Range	Lower Range
Apps	81.0	48094.0	2848.0	7896.0	-3496.0
Accept	72.0	26330.0	1820.0	5154.0	-2126.0
Enroll	35.0	6392.0	660.0	1892.0	-748.0
Top10perc	1.0	96.0	20.0	65.0	-15.0
Top25perc	9.0	100.0	28.0	111.0	-1.0
F.Undergrad	139.0	31643.0	3013.0	8524.5	-3527.5
P.Undergrad	1.0	21836.0	872.0	2275.0	-1213.0
Outstate	2340.0	21700.0	5605.0	21332.5	-1087.5
Room.Board	1780.0	8124.0	1453.0	7229.5	1417.5
Books	96.0	2340.0	130.0	795.0	275.0
Personal	250.0	6800.0	850.0	2975.0	-425.0
PhD	8.0	103.0	23.0	119.5	27.5
Terminal	24.0	100.0	21.0	123.5	39.5
S.F.Ratio	2.5	39.8	5.0	24.0	4.0
perc.alumni	0.0	64.0	18.0	58.0	-14.0
Expend	3186.0	56233.0	4079.0	16948.5	632.5
Grad.Rate	10.0	118.0	25.0	115.5	15.5

FIGURE 9: IQR, UPPER AND LOWER RANGE TABLE

Based on observations, we can see that of wide range of variability in values. These IQR, Upper and lower range shows about the presence of outlier. We have to perform the outlier treatment before the scaling. As per case study instruction, outlier treatment steps have to be skipped.

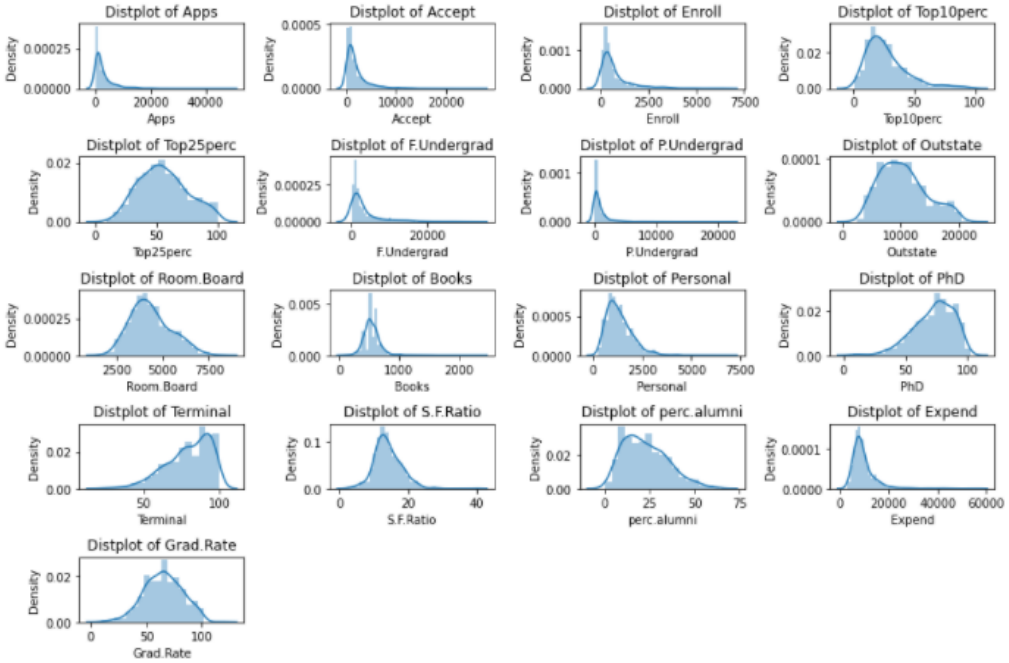

FIGURE 10: DISTRIBUTION PLOT FOR DATASET

Based on the Distplot, we can observe that for Apps, Top10perc, S.F. Ration, Expand, Accept, Enroll, F. Undergrad and P. Undergrad are left skewed. Similarly, PhD and Terminal are right skewed data spread. Grad Rate, Perc. Alumni, Top25Perc, Room board and Outstate all are normally bell curved distributed.

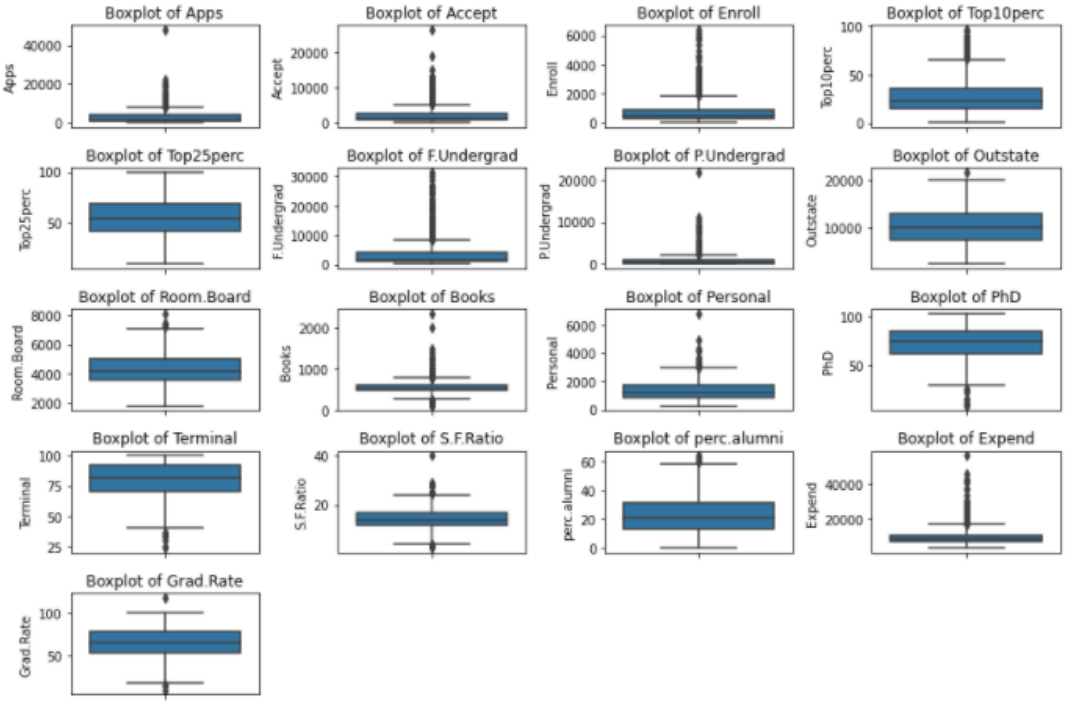

FIGURE 11: BOXPLOT FOR OUTLIER IDENTIFICATION ON VARIOUS VARIABLES

Based on observations, we can state that almost all the variable contains outliers. We need to do the outlier treatment before proceed further. But, as per instructions, Outlier treatment will only be done when say to do so.

Multivariate Analysis

Based on observations, we can state that variable pairs such as app/accept, accept/enrol, top 25 perc/top10 perc, PhD/ terminal are strongly correlation with each other.

Heatmap also exhibits the issue of multicollinearity in the dataset.

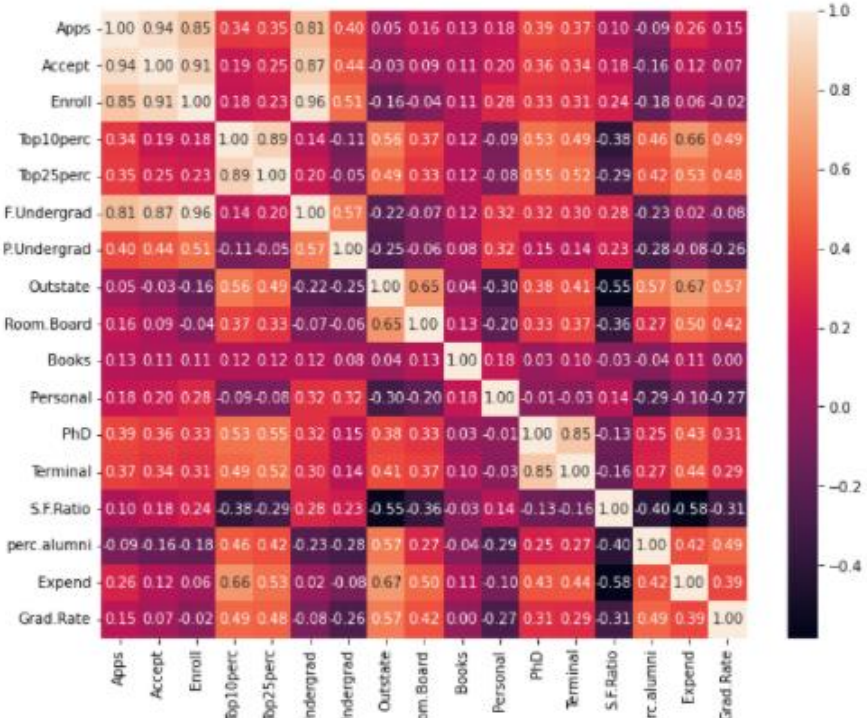

FIGURE 12: HEATMAP ANALYSIS FOR VARIOUS FEATURES

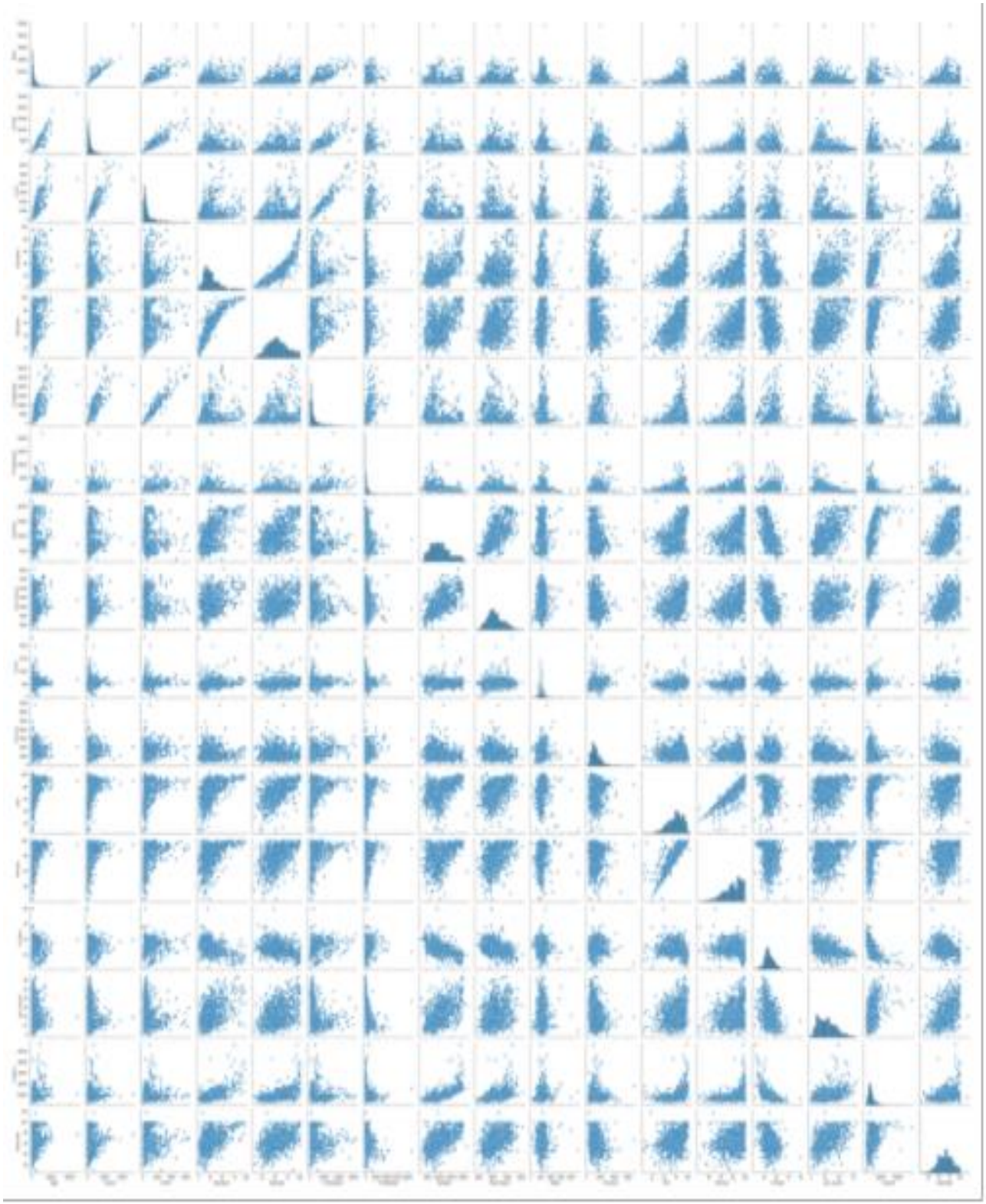

FIGURE 13: PAIRPLOT ANALYSIS FOR VARIOUS FEATURES

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

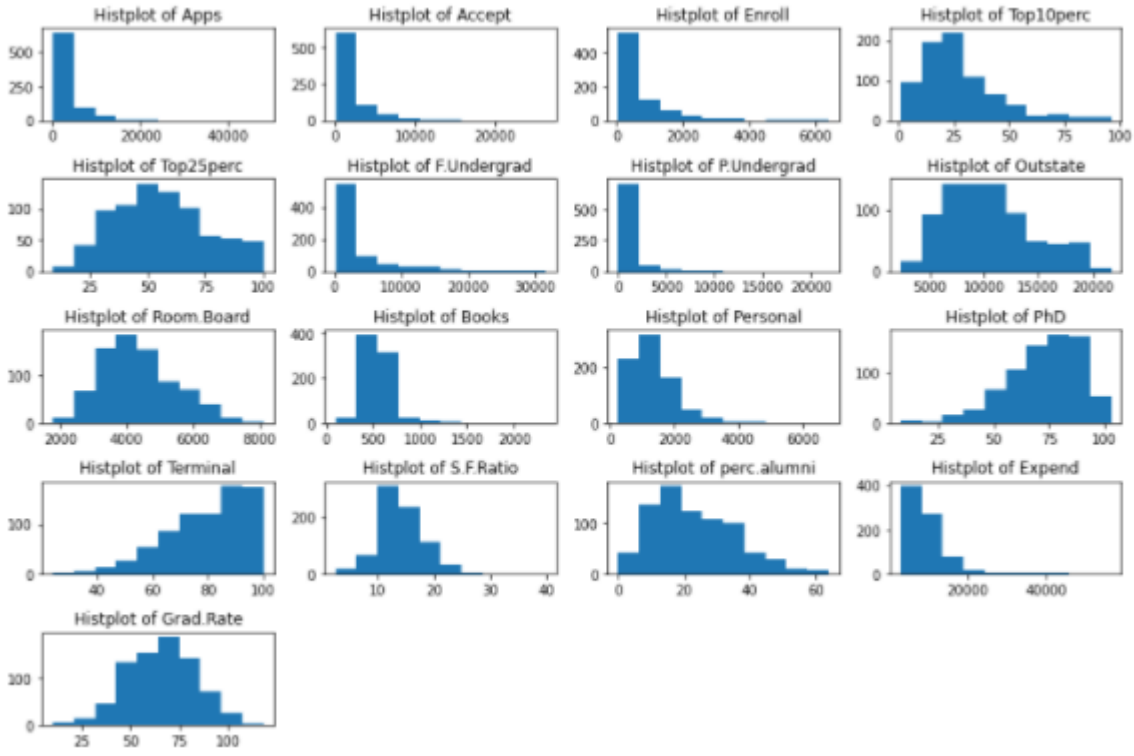

FIGURE 14: HISTOGRAM FOR VARIOUS FEATURE BEFORE SCALING

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.013776	-0.867574	-0.501910	-0.318252
-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	-0.477704	-0.544572	0.166110	-0.551262
-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	-0.300749	0.585935	-0.177290	-0.667767
-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	-1.615274	1.151188	1.792851	-0.376504
-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	-0.553542	-1.675079	0.241803	-2.939613
-0.624307	-0.628611	-0.669812	0.592287	0.313426	-0.623421	-0.535212	0.760947	-0.932970	-0.299280	-0.963753	-0.346878	-0.455567	-1.185526	-0.948325	0.012806	-0.609514
-0.684808	-0.685356	-0.729043	-0.598931	-0.545505	-0.677472	-0.410988	0.708713	1.243144	-0.299280	0.235515	1.062639	0.903786	-0.654660	0.262933	-0.153145	-0.143495
-0.285088	-0.121984	-0.313353	0.535563	0.616579	-0.434450	-0.541127	0.852479	0.427443	-0.602312	-0.725120	1.001356	1.379560	-0.098515	1.151188	0.350074	0.439030
-0.507700	-0.481644	-0.595505	0.138490	0.363952	-0.562562	-0.361036	1.282036	0.038754	-1.511408	-1.242385	0.388522	0.292077	-0.705218	0.020681	0.380160	0.846798
-0.625600	-0.620854	-0.654735	-0.372032	-0.596031	-0.598459	-0.510893	0.006798	-0.891911	0.670422	0.678885	-2.001529	-2.630532	-0.654660	-0.625323	-0.128233	-0.784272
-0.328266	-0.242415	-0.331661	0.535563	0.970257	-0.385763	-0.489860	1.519075	0.956645	-0.299280	-1.094595	0.572372	0.563948	-0.705218	0.666685	0.243720	0.439030
-0.090399	-0.048501	-0.318738	0.932636	1.071307	-0.411138	-0.533240	1.651399	0.075250	-0.905344	-1.094595	0.020822	0.767851	-1.059129	1.474190	0.392999	0.613788
-0.471245	-0.505730	-0.527659	0.592287	0.414477	-0.530173	-0.142823	-0.186714	0.390034	0.306784	-0.503435	-0.775861	0.292077	-0.199632	-0.140820	-0.329635	0.497283
-0.448492	-0.383258	-0.425352	0.932636	0.869206	-0.493864	-0.543756	0.530125	0.177441	-0.905344	-1.390175	0.388522	0.495980	0.305955	0.747436	-0.068061	0.147768
-0.648352	-0.696379	-0.670889	-0.258583	-0.494980	-0.491595	0.249565	-0.519514	-0.654683	0.609816	1.638042	-2.246662	-0.727438	-0.755777	0.262933	-0.293801	-0.609514
-0.408934	-0.377951	-0.603044	-1.052728	-1.707589	-0.553279	-0.373524	-0.432956	0.385472	-0.602312	0.087725	0.327239	0.292077	0.154279	-0.302321	-0.441738	0.206020
0.336210	-0.419183	-0.389814	3.144898	2.031290	-0.434656	-0.558873	2.317995	0.859929	0.670422	0.380349	1.246489	1.243624	-1.438319	3.250701	2.254295	2.011847

FIGURE 15: SCALED DATASET

Our dataset has 18 attributes. Initially, data range variation is very much. It is important to normalize data before performing PCA. The Standard scaler assumes that data is normally distributed within each feature and scaled data distribution is now centered around 0, with a SD of 1.

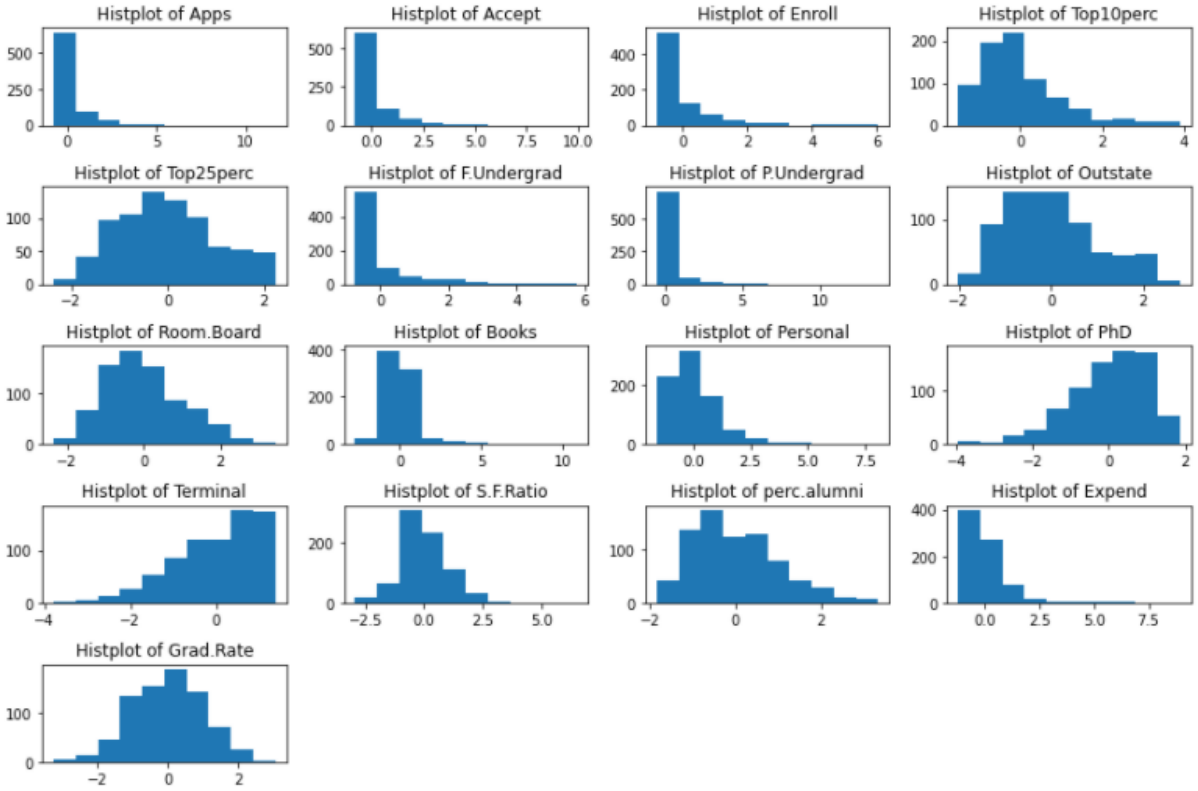

FIGURE 16: HISTOGRAM FOR VARIOUS FEATURE AFTER SCALING

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

The term covariance and correlation both sync with each other. In one way where covariance indicates the direction of linear relationship between variables on the other hand correlation denotes strength and direction of linear relationship between two variables. Scaling affects the covariance. Below metrics shown about the correlations between variables and their relationship strength.

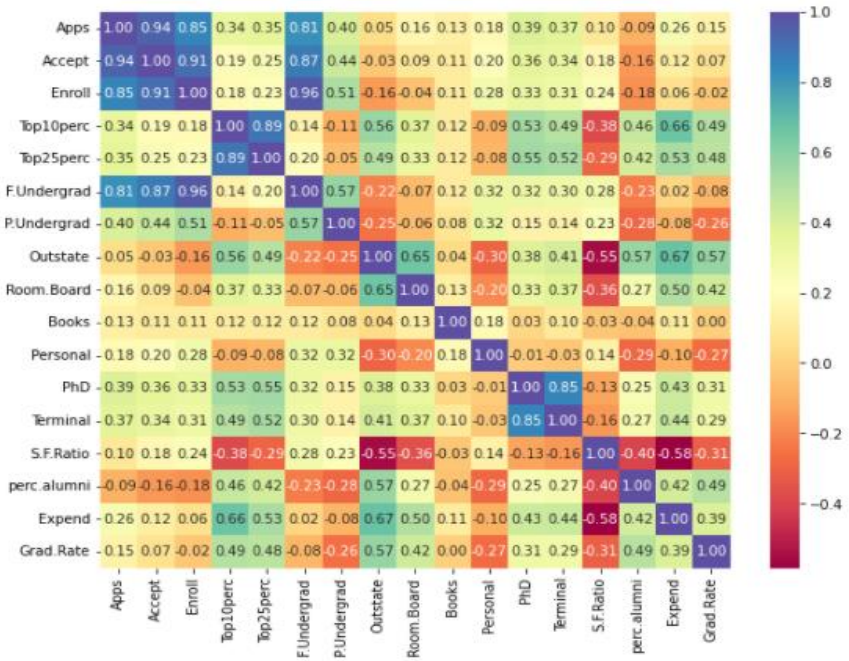

FIGURE 17: CORRELATIONSHIP METRICS BASED ON SCALED DATASET

Strong Negative Correlation:

Expend	S.F.Ratio	-0.583832
S.F.Ratio	Outstate	-0.554821
perc.alumni	S.F.Ratio	-0.402929
S.F.Ratio	Top10perc	-0.384875
	Room.Board	-0.362628
Grad.Rate	S.F.Ratio	-0.306710

Strong Positive Correlation:

F.Undergrad	Apps	0.814491
Enroll	Apps	0.846822
Terminal	PhD	0.849587
F.Undergrad	Accept	0.874223
Top25perc	Top10perc	0.891995
Enroll	Accept	0.911637
Accept	Apps	0.943451
F.Undergrad	Enroll	0.964640

Based on observation of scaled dataset, we have figured out some strong positive and negative correlated pairs.

2.4 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]

Below figure reflects the result of boxplot on raw dataset and scaled data. We can see that almost all variables have outliers' present in the result. There is no effect of scaling in outlier presence. Scaling helps in to shrinks the range of the feature. It has no effect on outlier reduction.

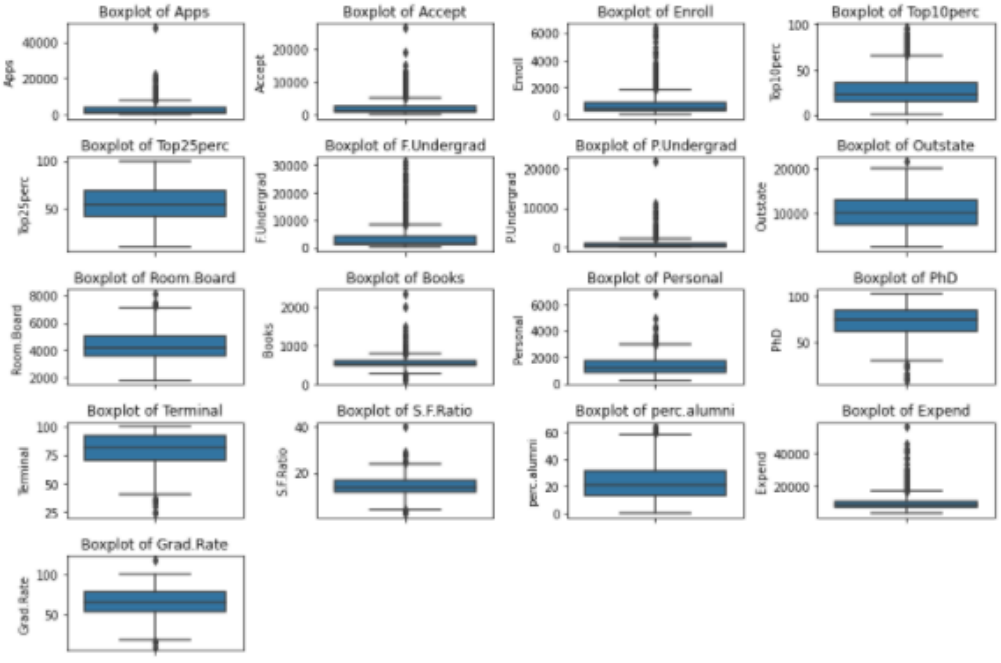

FIGURE 18: OUTLIER STATE BEFORE SCALING

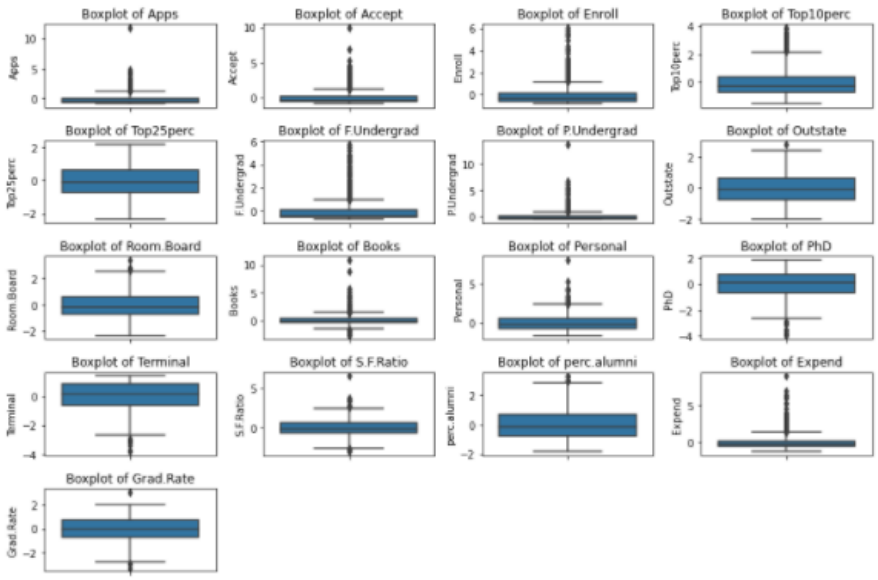

FIGURE 19: OUTLIER STATE AFTER SCALING

2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

Eigen Values:

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.31344588, 0.22061096, 0.16779415, 0.1439785 , 0.08802464,
       0.03672545, 0.02302787])
```

Eigen Vectors:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.358970
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-0.085697	0.029590	-0.404318	0.609651
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	0.151742	-0.617275	0.051868	0.080348
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-0.056373	0.009916	0.560363	-0.414705
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	0.019286	0.020952	-0.052731	0.009018
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-0.034012	0.038354	0.101595	0.050900
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-0.058429	0.003402	-0.025929	0.001146
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011438	-0.066849	-0.009439	0.002883	0.000773
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	0.027529	-0.003090	-0.012890	-0.001114
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-0.691126	-0.112056	0.029808	0.013813
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	0.671009	0.158910	-0.027076	0.006209
S.F.Ratio	-0.176958	0.246665	-0.289648	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	0.041374	-0.020899	-0.021248	-0.002222
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-0.027154	-0.008418	0.003334	-0.019187
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	0.073123	-0.227742	-0.043880	-0.035310
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	0.036477	-0.003394	-0.005008	-0.013071

FIGURE 20: PCA ON ORIGINAL DATAFRAME FEATURES

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

Based on PCA analysis, we can see that below scree plot shows that 33 % variability of data comes from first PCA. Till PCA 8, variability spread reduces up to 5%.

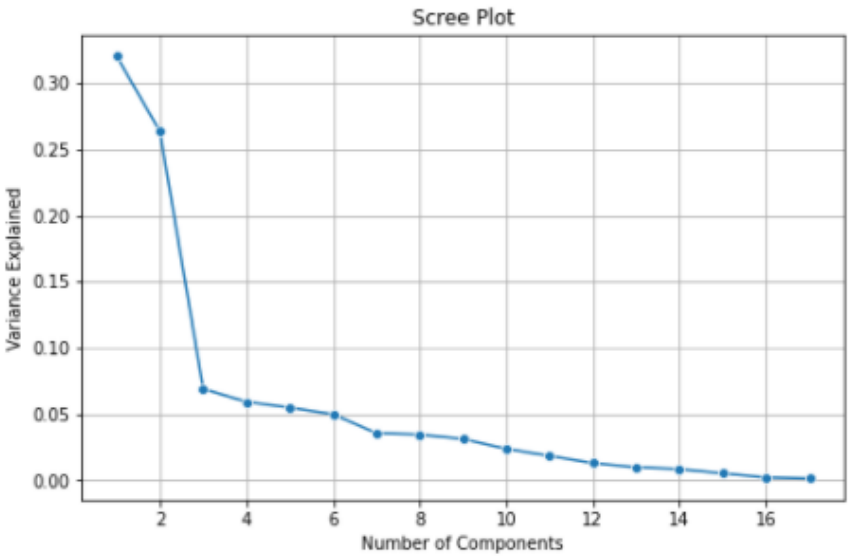

FIGURE 21: SCREE PLOT

All the PCA into the data frame with the original features has been captured in subsequent figures. Based on scree plot and other variance ratio analysis, first 8 PCA are very much essentials for the analysis. Finally, 9 PCA correlation matrices has created which shows the independent ability of the vectors.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.358970
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-0.085697	0.029590	-0.404318	0.609651
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	0.151742	-0.617275	0.061868	0.080348
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-0.056373	0.009916	0.560363	-0.414705
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	0.019286	0.020952	-0.052731	0.009018
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-0.034012	0.038354	0.101595	0.050900
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-0.058429	0.003402	-0.025929	0.001146
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011498	-0.066849	-0.009439	0.002883	0.000773
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	0.027529	-0.003090	-0.012890	-0.001114
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-0.691126	-0.112056	0.029808	0.013813
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	0.671009	0.158910	-0.027076	0.006209
S.F.Ratio	-0.176958	0.246665	-0.269648	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	0.041374	-0.020899	-0.021248	-0.002222
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-0.027154	-0.008418	0.003334	-0.019187
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	0.073123	-0.227742	-0.043880	-0.035310
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	0.036477	-0.003394	-0.005008	-0.013071

FIGURE 22: PRINCIPAL COMPONENTS ON DATAFRAME FEATURES

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904

FIGURE 23: PRINCIPAL COMPONENTS AFTER DIMENSION REDUCTION

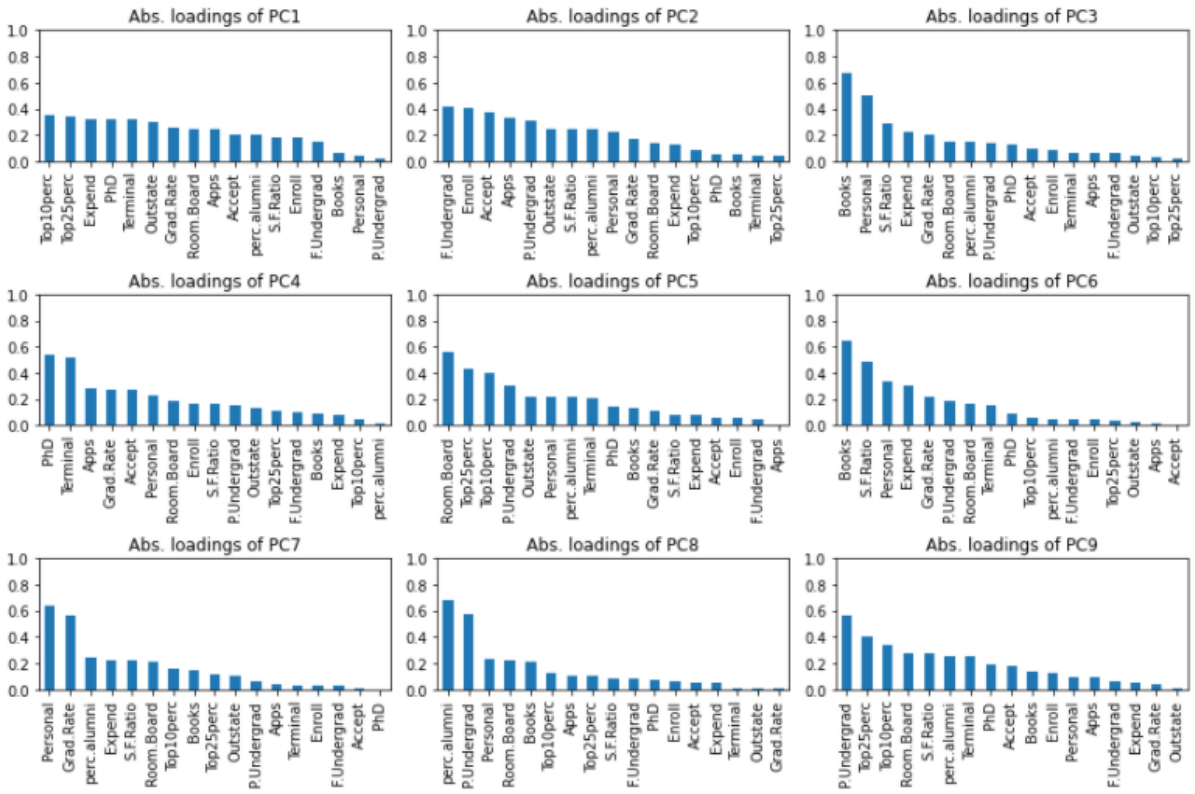

FIGURE 24: LOADING OF EACH PRINCIPAL COMPONENTS

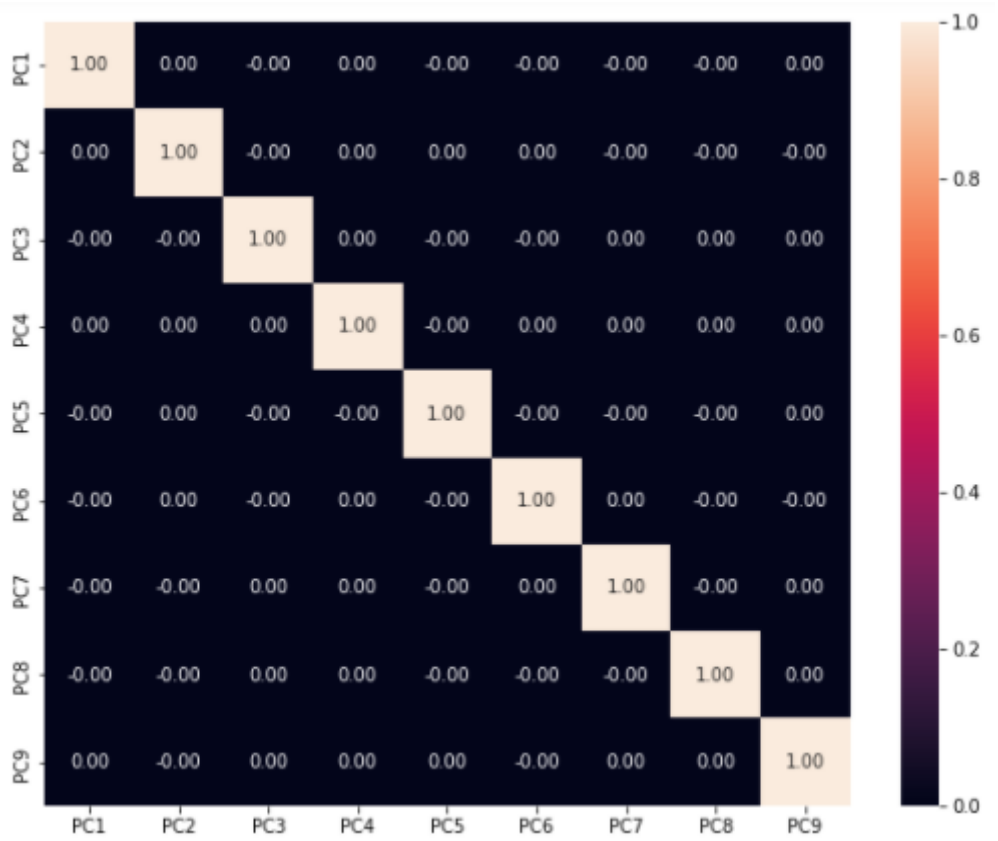

FIGURE 25: CORRELATION METRICS AFTER PCA

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

Equation on dataset variables X and corresponding weights W reflects as shown below. The first principal component PC1 isshown by the linear combination of the original variables.

$$PC\_1 = w\_\{17\}X\_1 + w\_\{16\}X\_2 + \ldots + w\_\{1p\}X\_p$$

The explicit form of the PC1 is as below:

```
array([ 0.25,  0.21,  0.18,  0.35,  0.34,  0.15,  0.03,  0.29,  0.25,
        0.06, -0.04,  0.32,  0.32, -0.18,  0.21,  0.32,  0.25])
```

2.8 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?

```
array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
       0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,
       0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,
       0.99864716, 1.        ])
```

Below plot shows that how much variance comes from a various PCA. As per plot, 1 PC covers 33%, 2nd almost 57% and list goes on. From the analysis and plot, almost 9 PCA covers approx. 90% of range of dataset. Eigen vectors determines the direction or axes along which the linear transformation acts or shrinking the input vector.

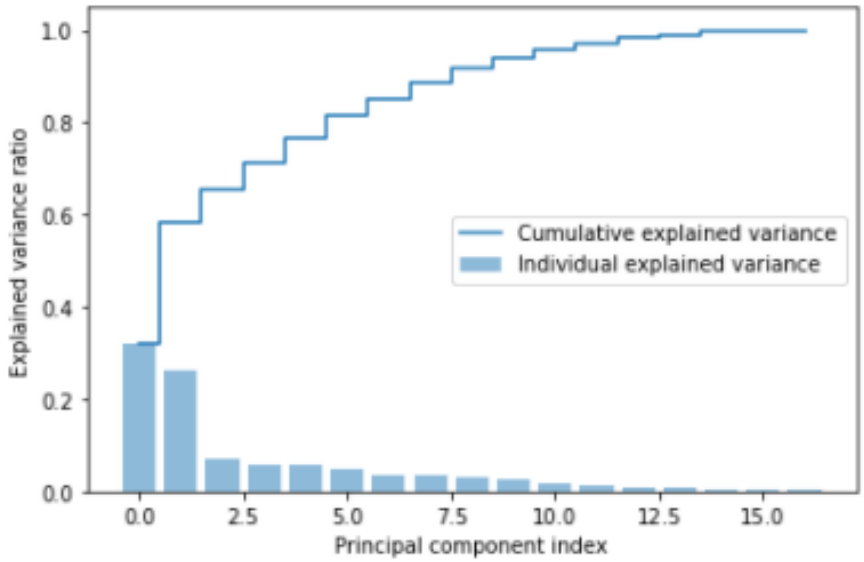

FIGURE 26: EXPLAINED VARIANCE RATIO PLOT

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: **Write Interpretations of the Principal Components Obtained]**

Business implication of using the Principal Component Analysis for this case study is very essential. PCA is used for dimension reduction. Problem has started with 17 variables which brings down to 9 with almost 90% data coverage. Due to this step, we have reduced the model complexity and computation usage of the system. Better the EDA along with PCA brings the accuracy and clean analysis. First principal component can reflect the direction which max. the variance of project data. One principal component is orthogonal to other PC to cover up the variance of dataset.