PROJECT: ADVANCED STATISTICS

ADS

greatlearning
Learning for Life

Problem 1 0

1.1State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually……………………………………………………..4

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…..........4

1.3 Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.....4

1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result...0

1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot........4

1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result?...............................4

1.7 Explain the business implications of performing ANOVA for this particular case study.2

Problem 2……………………………………………………………………………………………0

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?...4

2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling……………………………………………………………………………………2

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

2.4 Check the dataset for outliers before and after scaling. What insight do you Derive here?...3

2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]…………..4

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features……………………………………………………………..4

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]…………………………………………………4

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?....4

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]……………………………………………………….4

Quality of Business Report (Please refer to the Evaluation Guidelines for Business report checklist. Marks in this criteria are at the moderator's discretion)……………………..6

Please reflect on all that you learnt and fill this reflection report - https://forms.gle/2DF5i1W7sUXXmsaw5...0

1.1. 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

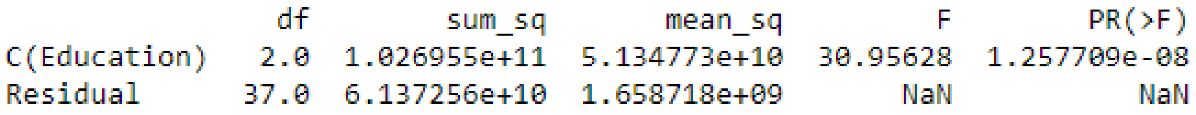
- H0: Salary depend on education qualification
- Ha: Salary does not depend on education
- Confidence level = 0.05

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

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

Solution –

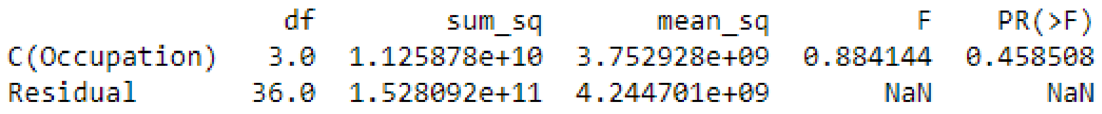


```
                 df         sum_sq       mean_sq          F        PR(>F)
C(Education)    2.0   1.026955e+11  5.134773e+10   30.95628  1.257709e-08
Residual       37.0   6.137256e+10  1.658718e+09        NaN           NaN
```

Since the p-value is less than Alpha we reject Null Hypothesis (H0) for Education.

1.3. Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Solution –



```
                 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
```

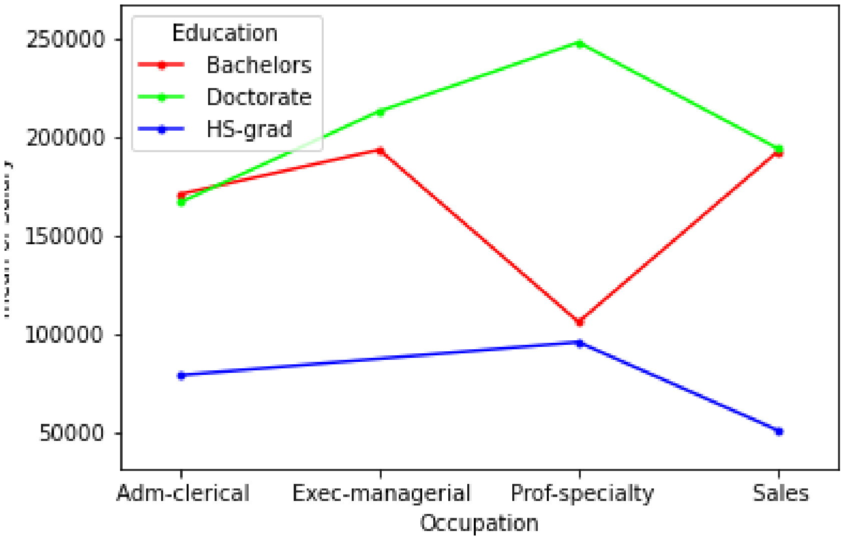
Since the p-value is greater than Alpha we cannot reject Null Hypothesis (H0) for Occupation.

1.4. If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result

Solution – NA

1.5. What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

Solution –



Observation –

- ✓ From above plot we can make out the interaction as followed:
  - Adm-clerical job with Bachelors & Doctorate is fairly good.
  - Sales Job with Bachelors & Doctorates is also good.
  - Exec-managerial Job role has no interactions with any other educational background.
- ✓ From Above plot also we can figure out below observations –
  - Doctorates are into higher salary brackets & mostly Prof-specialty roles or Exec-managerial or in Sales profiles, very few are doing Adm-clerical Jobs.
  - Bachelors fall in mid income range and found mostly working as an Exec-managers, Adm-clerical or into sales but very few are found in Prof-specialty profile.
  - HS-grads are in low income brackets, mostly doing Prof-specialty or Adam-clerical work and few are doing Sales but hardly any in Exec-managerial role.

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

Solution –

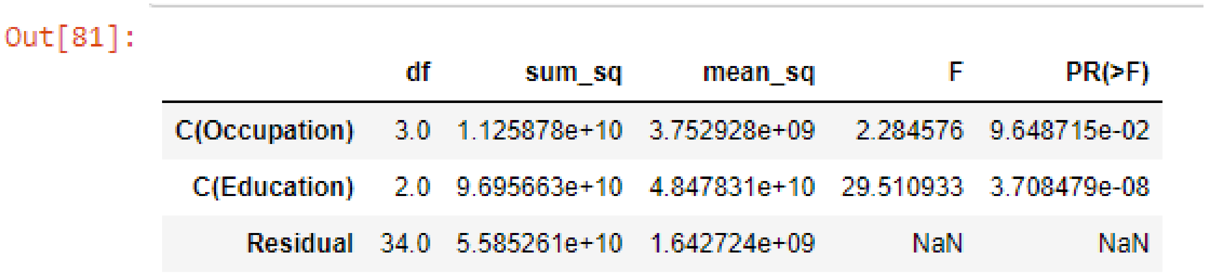
The Null and Alternative Hypothesis for the Two Way ANOVA for each Occupation type & Education

Level are:

H0: The mean Salary variable for each Occupation type & Education level are Equal
Ha: For at least one of the means of Salary for type of Occupation and Education Level are not equal

Where Alpha=0.05

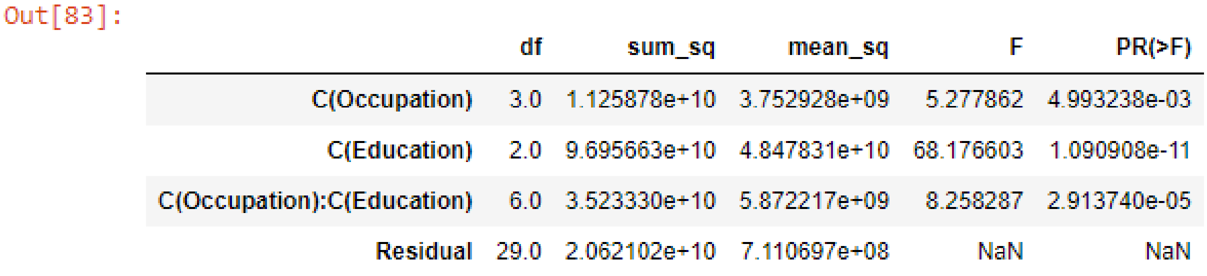
- If the p-value<0.05, then we reject the Null Hypothesis.
- If the p-value>0.05, then we fail to reject the null Hypothesis.
-



Out[81]:

	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

Still, we can see that there is some sort of interaction between the two treatments. So, we will introduce a new term while performing the Two Way ANOVA



Out[83]:

	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(Occupation):C(Education)	6.0	3.523330e+10	5.872217e+09	8.258287	2.913740e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Due to the inclusion of the interaction effect term, we can see a slight change in the p-value of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms.

And we see that the p-value of the interaction effect term of ‘’Education” and “Occupation” suggests that the Null Hypothesis is rejected in this case.

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

Solution –

ANOVA stands for “analysis of variance” and is used in statistics when you are testing a hypothesis to understand how different groups respond to each other by making connections between independent and dependent variables. ANOVA is a statistical test that compares the means of groups in order to determine if there is a difference between them. It is used when more than two group means are compared. For two group means, we can do t-test.
 ANOVA is used in a business context to help manage income /salary by comparing your education to occupation here in this case to help manage revenue income (salary).
 ANOVA can also be used to forecast Salary trends by analyzing patterns in data to better understand the future hike of Salary.
 It is also a widely used statistical technique for comparing the relationship between factors that cause a rise in Salary, assuming this report is for HR department or HR consulting firm. Some of the key takeaways as below:

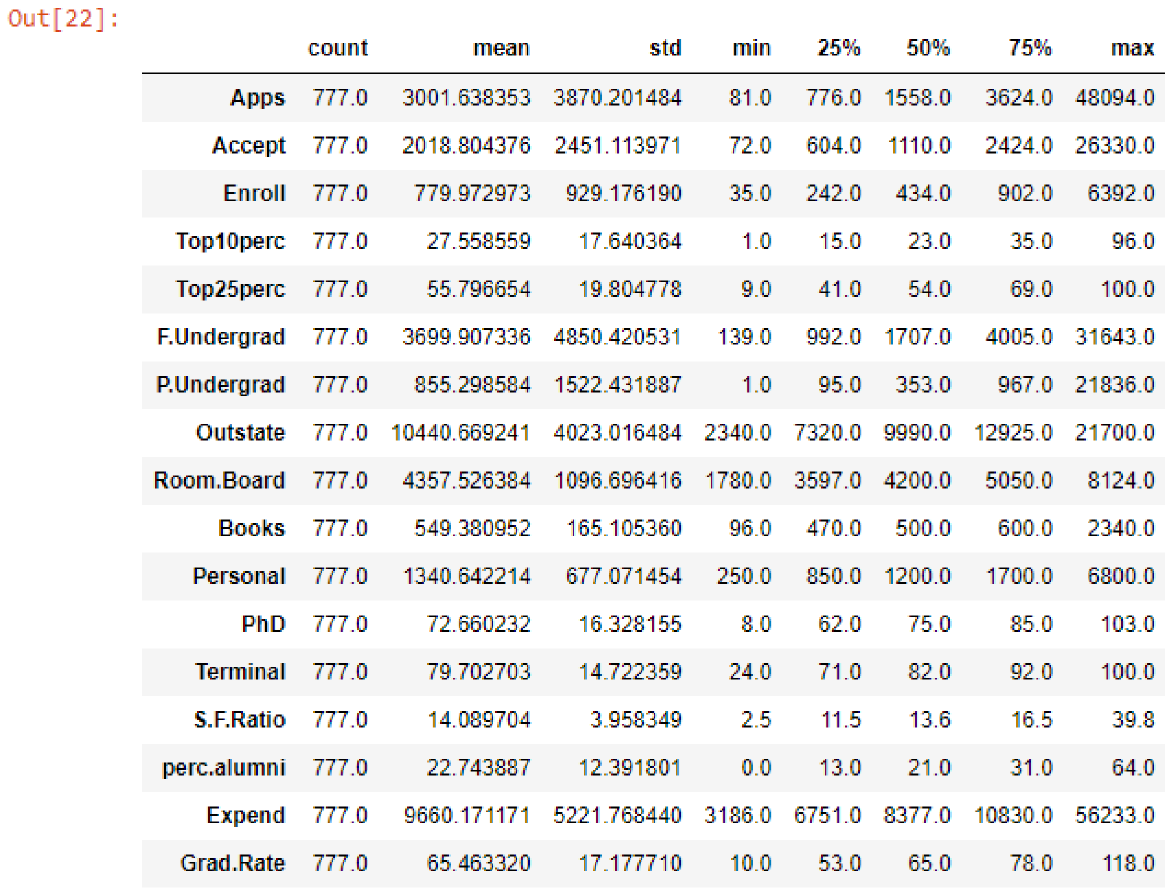
i. As the Education level upgrades Salary increases. On an average Doctorate earns higher salaryth an Bachelors and HS-Grads. However, it might be possibility that being Doctorate may not necessarily mean significant high salary than HS-Grad or Bachelors employees. So that means Doctorates are suitable for all job role or not always preferred above other education levels, maybe they can be considered some times as over qualified for certain job roles

ii. Though there is lesser significance of Occupation than education on Salary but at certain levels it impacts Salary.

iii. We must also take note of that high salaries are offered to Bachelor's degree holders than Doctorates for few occupations. So, we can say that there are some shortcomings of dataset provided which reduces accuracy of the test and analysis done, as there can be few more other important variables which can impact salary such as years of experience, specialization, industry/domain etc.

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

Solution –



Out[22]:

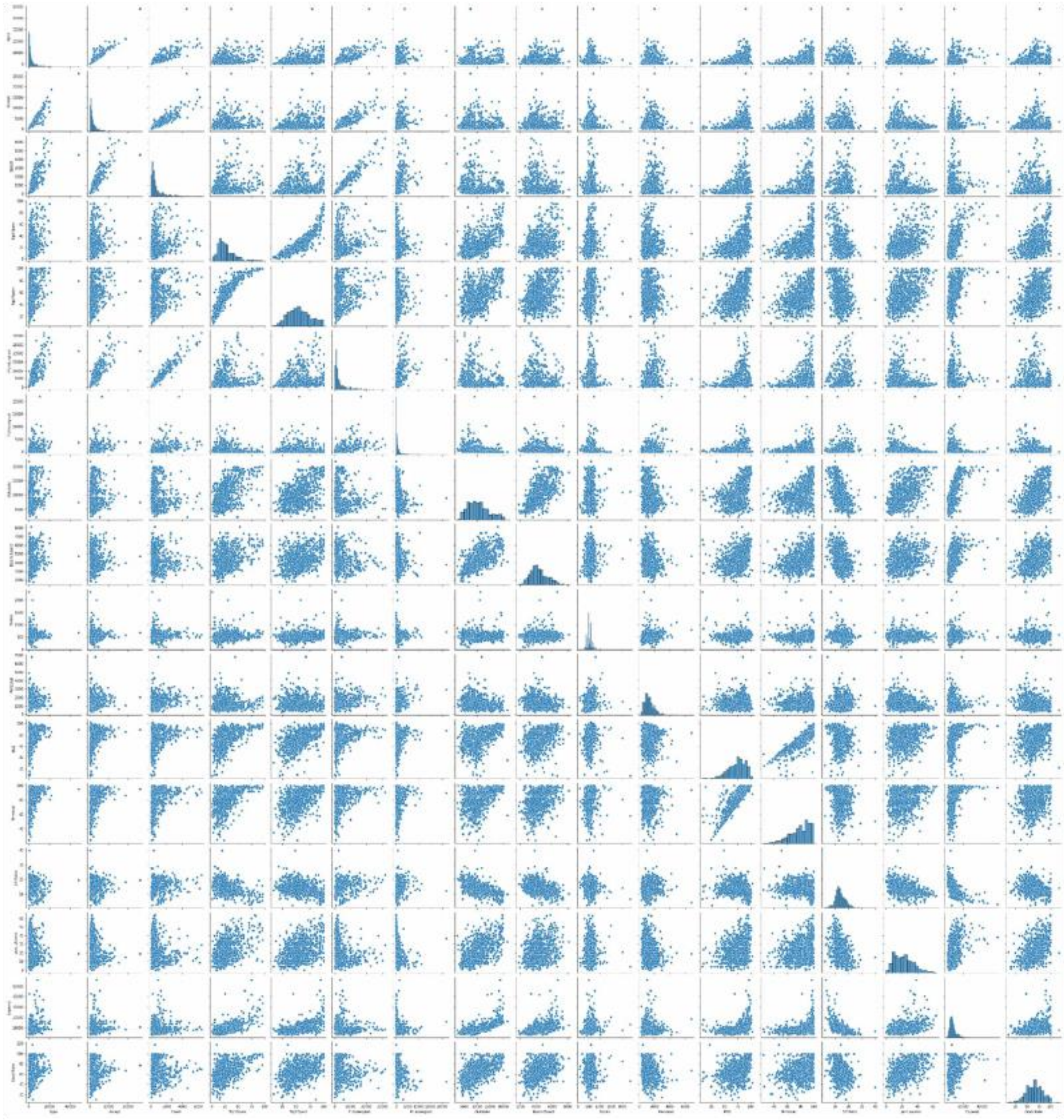
	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	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

Observations:

□ Data consists of 777 Universities with 18 Variables but not a single categorical variable present.
□ Index of columns : 'Name', 'Apps', 'Accept', 'Enroll', 'Top10perc', 'Top25perc', 'F.Undergrad','P.Undergrad', 'Outstate', 'Room.Board', 'Books', 'Personal', 'PhD', 'Terminal', 'S.F.Ratio','perc.alumni', 'Expend', 'Grad. Rate'.
□ Perc.alumni have minimum values as 0. Needs to be cleaned
□ There are no missing values in the data.
□ We have 1 categorical field Name field needs cleanup.

- Very few students fall under topper students with Top 10% and 25%
- No duplicate records
- P.Undergrad has a minimum value as 1 and maximum value as 21836.. This has to cleaned.
- Accept field needs cleanup, has a minimum value as 72 and maximum value as 26330.
- Apps field needs cleanup, has a minimum value as 81 and maximum value as 48094.
- Books has a minimum value as 96 and maximum value as 2340.. This has to cleaned.
- Enroll has a minimum value as 35 and maximum value as 6392.. This has to cleaned.
- FUndergrad field needs cleanup, has a minimum value as 139 and maximum value as 31643.
- From Quick Summary table we found that max % of Graduated students are 118% which needs to be rectified as it shouldn’t go beyond 100. So the error was found on row 95 for Cazenovia College has to be corrected using Median value
- From Quick Summary table we found that max % of Phd students are 103% which needs to be rectified as it shouldn’t go beyond 100. So the error was found on row 582 and has to be corrected using Median value.

Multivariate Analysis –

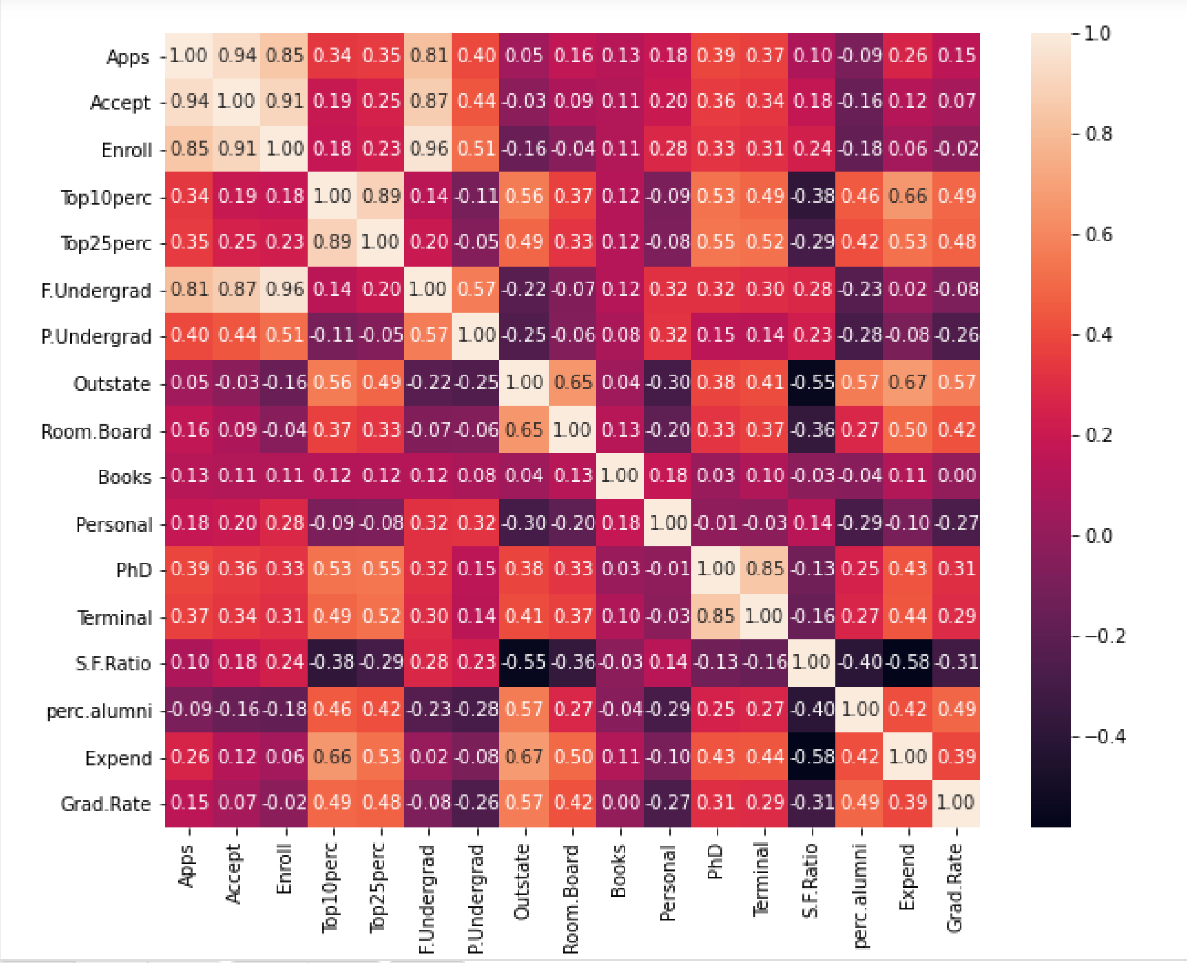


Observations –

Few pairs have very high correlation –

- Application & Acceptance
- Students from top 10% schools and from top 25% schools.
- Students from top 10% schools and Graduate rate
- Enrollment and full name undergrad students
- PHD faculties and Terminal

Below Heatmap exhibits multicollinearity issue as significant number of high correlation variables pairs/features.



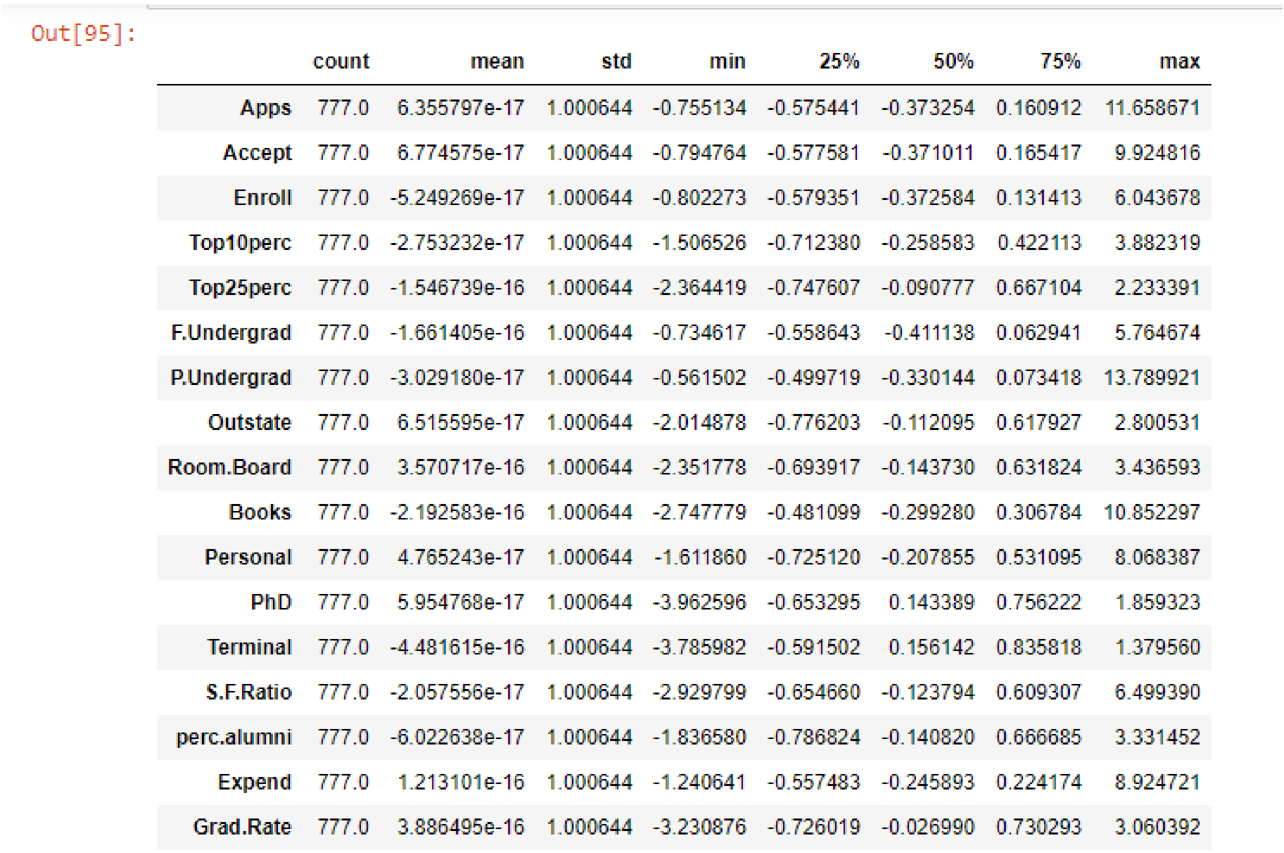
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37	-0.36	0.27	0.50	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.37	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
Grad.Rate	0.15	0.07	-0.02	0.49	0.48	-0.08	-0.26	0.57	0.42	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

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

Solution –

- Our Dataset has 18 attributes initially hence we get 18 principal components.
- Once we get the amount of variance explained by each principal component we can decide how may components we need for our model based on the amount of information we want to retain.
- Hence, it is necessary to normalize data before performing PCA.
- Scaling of Data can be done using Z-score method –

Statistical Description of Scaled Dataset is as follows –



Out[95]:

	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

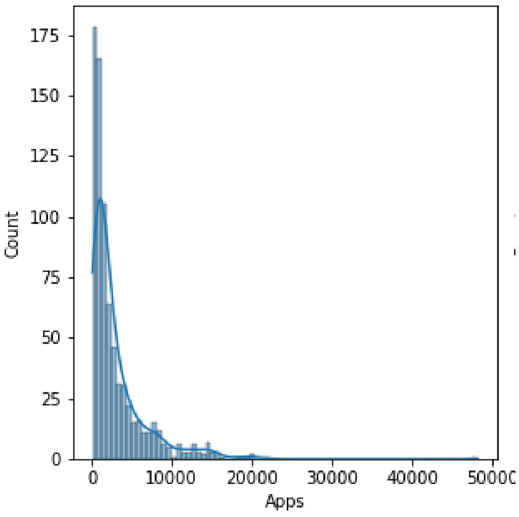
Observations –

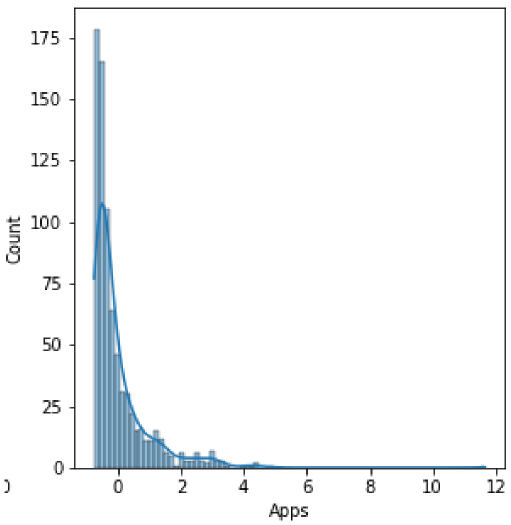
- After Scaling Standard deviation is 1.0 for all variables.
- Post Scaling Q1 (25%) value and min values difference is lesser than original dataset in most of the variables.

<u>Example after doing scaling -</u>

Before Scaling -

After Scaling -

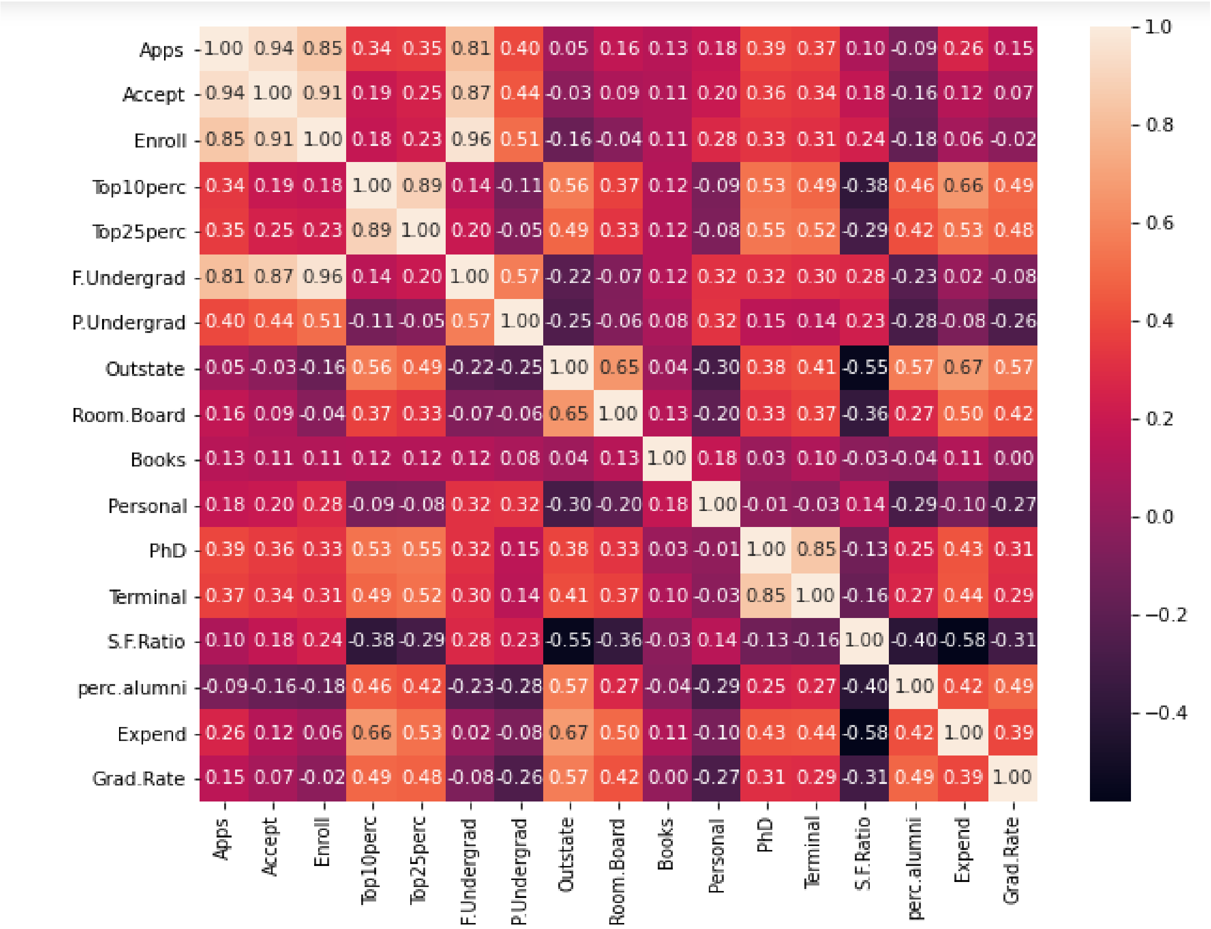




2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]

Solution –

□ Both the terms, Covariance and Corelation matrices measure the relationship and the dependency between two variables.
□ "Covariance" indicates the direction of the linear relationship between variables.
□ "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables.
□ Correlation refers to the scaled form of covariance. Covariance is
Affected by the change in scale.
□ Covariance indicates the direction of the linear relationship between variables. Correlation on the other hand measures both the strength and direction of the linear relationship between two variables.



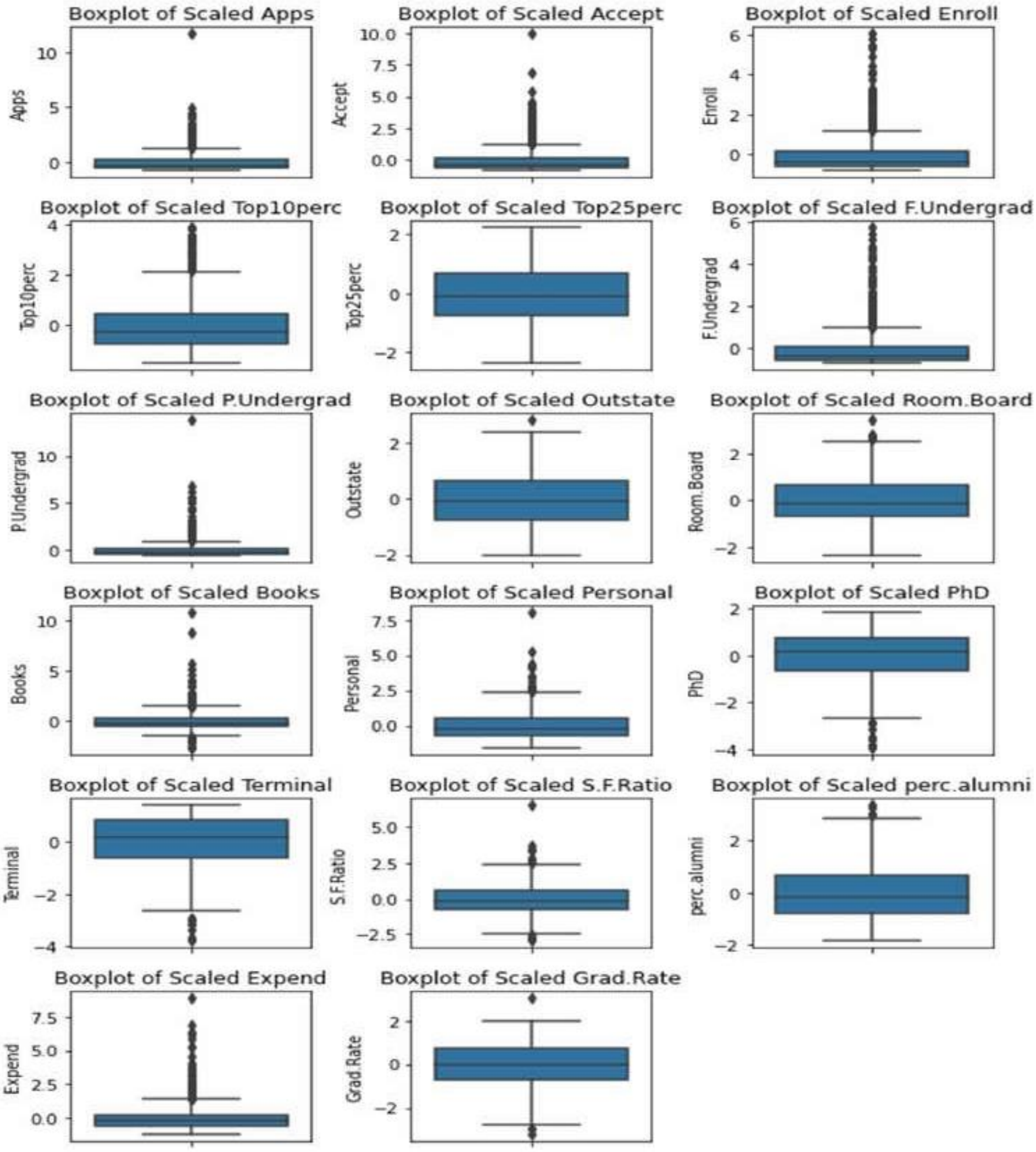
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37	-0.36	0.27	0.50	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.37	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
Grad.Rate	0.15	0.07	-0.02	0.49	0.48	-0.08	-0.26	0.57	0.42	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

Observations –

- Highest correlation is seen among:
  - Enroll variable with F.Undergrad
  - Enroll with Accept
  - Apps with Accept
- Least Correlations observed among
  - SF Ratio variable with Expend, Outstate, Grad Rate, perc.alumni, Room board and Top 10perc.

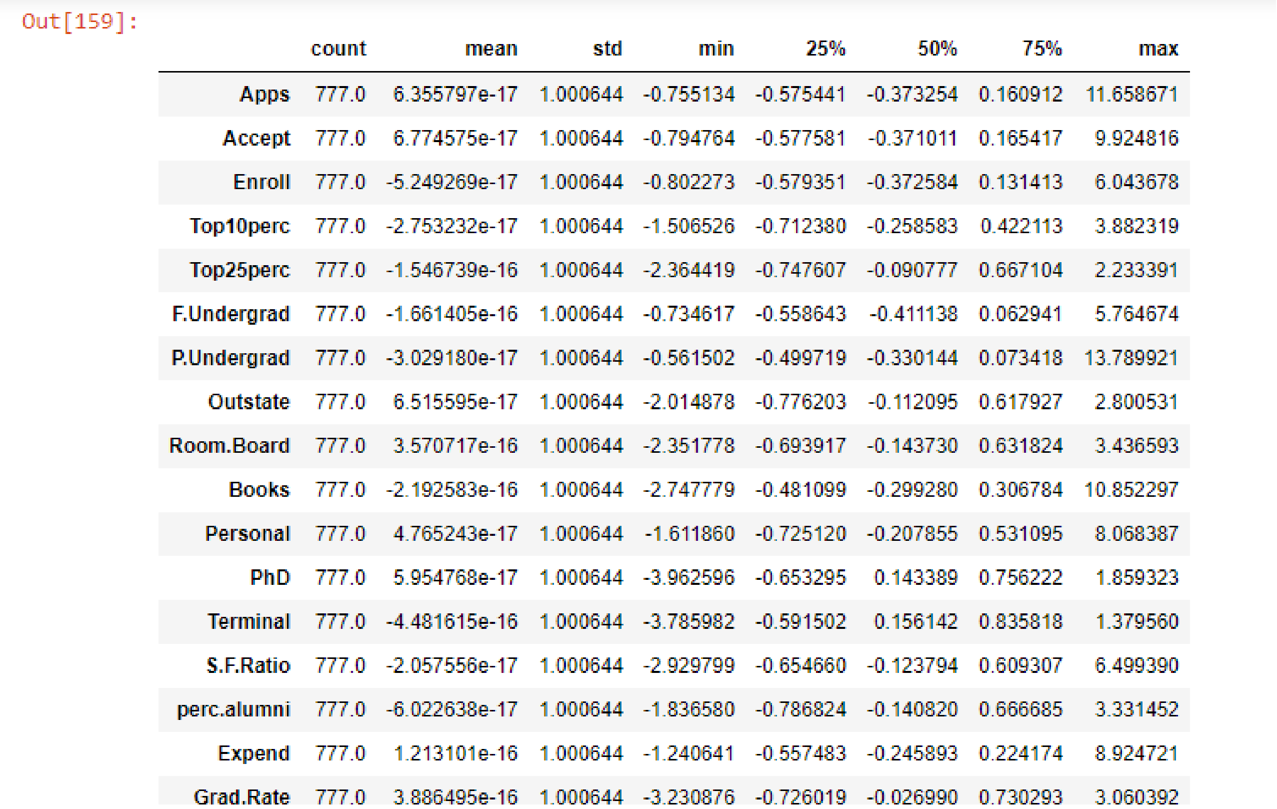
2.4 Check the dataset for outliers before and after scaling. What insight do you Derive here?

Solution –



The scaling shrinks the range of the feature values as shown in the left figure below. However, the outliers have an influence when computing the empirical mean and standard deviation. .StandardScaler therefore cannot guarantee balanced feature scales in the presence of outliers.

Scaled Data Summary



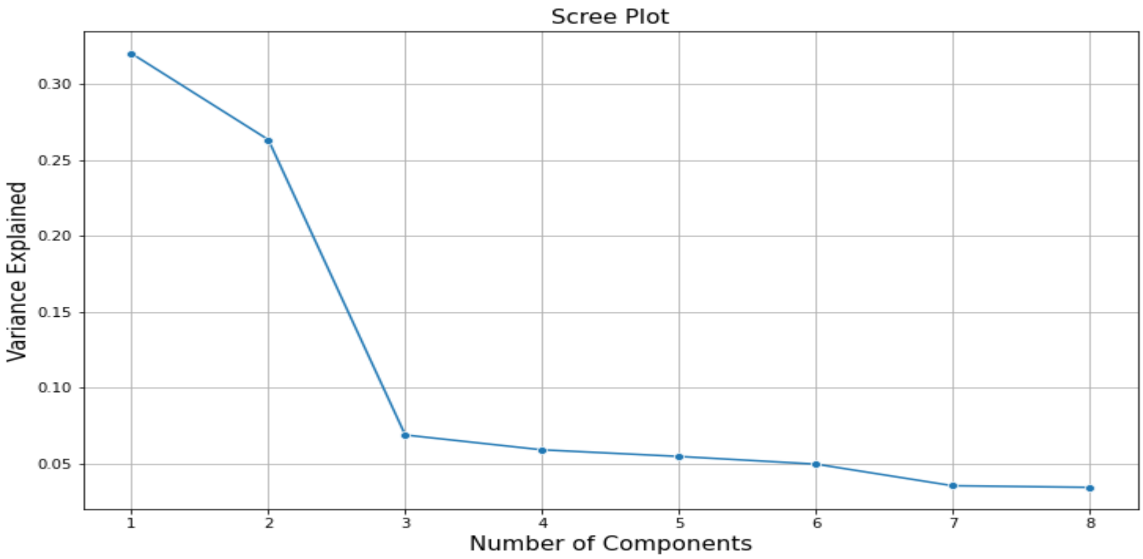
Out[159]:

	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

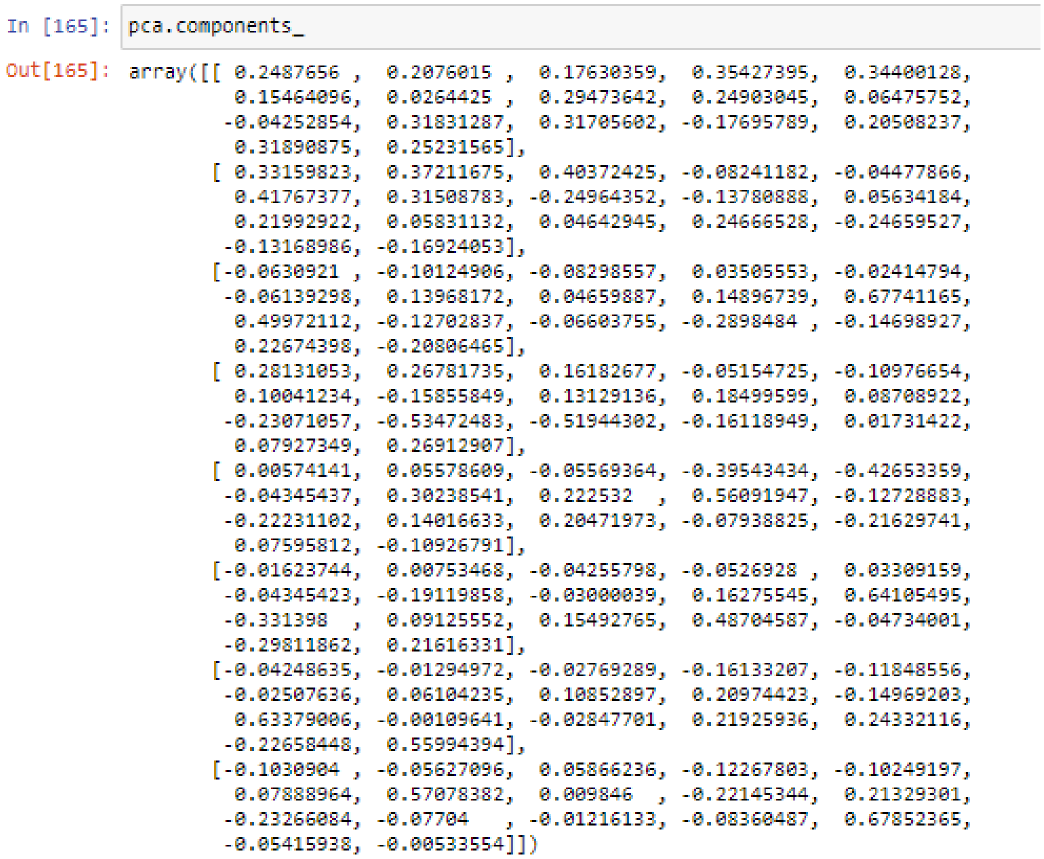
So, even if there are outliers in the data, they will not be affected by standardization.

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

Scree Plot showing distribution of PCs -



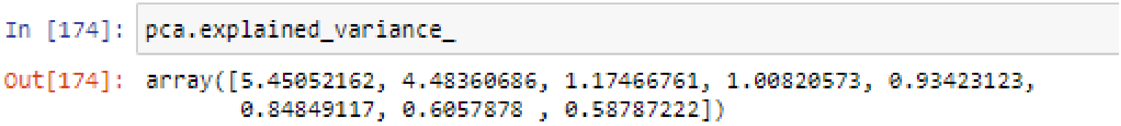
Eigen Vectors –



```
In [165]: pca.components_

Out[165]: 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, -0.08298557,  0.03505553, -0.02414794,
                  -0.06139298,  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,
                  -0.04345437,  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, -0.04255798, -0.0526928 ,  0.03309159,
                  -0.04345423, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
                  -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
                  -0.29811862,  0.21616331],
                 [-0.04248635, -0.01294972, -0.02769289, -0.16133207, -0.11848556,
                  -0.02507636,  0.06104235,  0.10852897,  0.20974423, -0.14969203,
                   0.63379006, -0.00109641, -0.02847701,  0.21925936,  0.24332116,
                  -0.22658448,  0.55994394],
                 [-0.1030904 , -0.05627096,  0.05866236, -0.12267803, -0.10249197,
                   0.07888964,  0.57078382,  0.009846  , -0.22145344,  0.21329301,
                  -0.23266084, -0.07704   , -0.01216133, -0.08360487,  0.67852365,
                  -0.05415938, -0.00533554]])
```

Eigen Values –

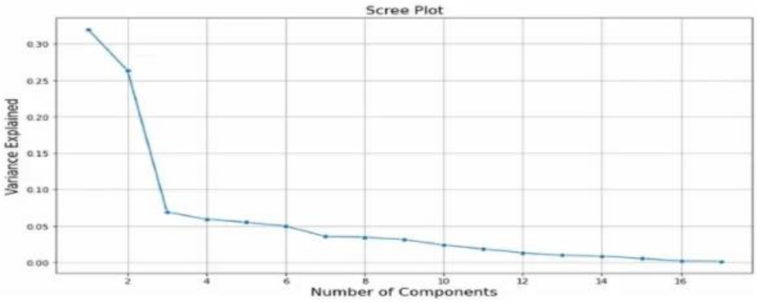


```
In [174]: pca.explained_variance_

Out[174]: array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
                 0.84849117, 0.6057878 , 0.58787222])
```

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

In below table we can see that first PC or Array explains 33.12% variance in our dataset, while first seven features captures 70.12% variance.



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]

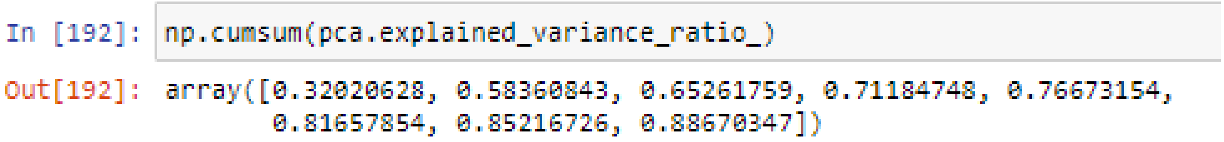
In PCA, given a mean centered dataset X with n sample and p variables, the first principal component PC1 is given by the linear combination of the original variables X_1, X_2, …, X_p
PC_1 = w_{17}X_1 + w_{16}X_2 + … + w_{1p}X_p

The first principal component PC1 represents the component that retains the maximum variance of the data. w1 corresponds to an
Eigenvector of the covariance matrix

The explicit form of the PC1 is as below:

[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],

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?



```
In [192]: np.cumsum(pca.explained_variance_ratio_)

Out[192]: array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
                 0.81657854, 0.85216726, 0.88670347])
```

Observations:
□ The above figure shows how much of the variance are explained, by how many principle components.
□ In the above figure we see that, the 1st PC explains variance 32.02%, 2nd PC explains 58.36% and so on.
□ Effectively we can get material variance explained (i.e. 90%) by analyzing 8 Principle components instead all of the 17 variables (attributes) in the dataset.

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]

The Business implication of using the Principal Component Analysis:

- PCA is used in exploratory data analysis and for making predictive models, can be done only on continuous variables.
- PCA used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data’s variation as possible. In this case we can reduce dimensions from 17 to 8 & that explains over 90% variances.
- The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data.
- The i th principal component can be taken as a direction orthogonal (i.e. at 90 degrees to one another) to the first i-1 principal components that maximizes the variance of the projected data.