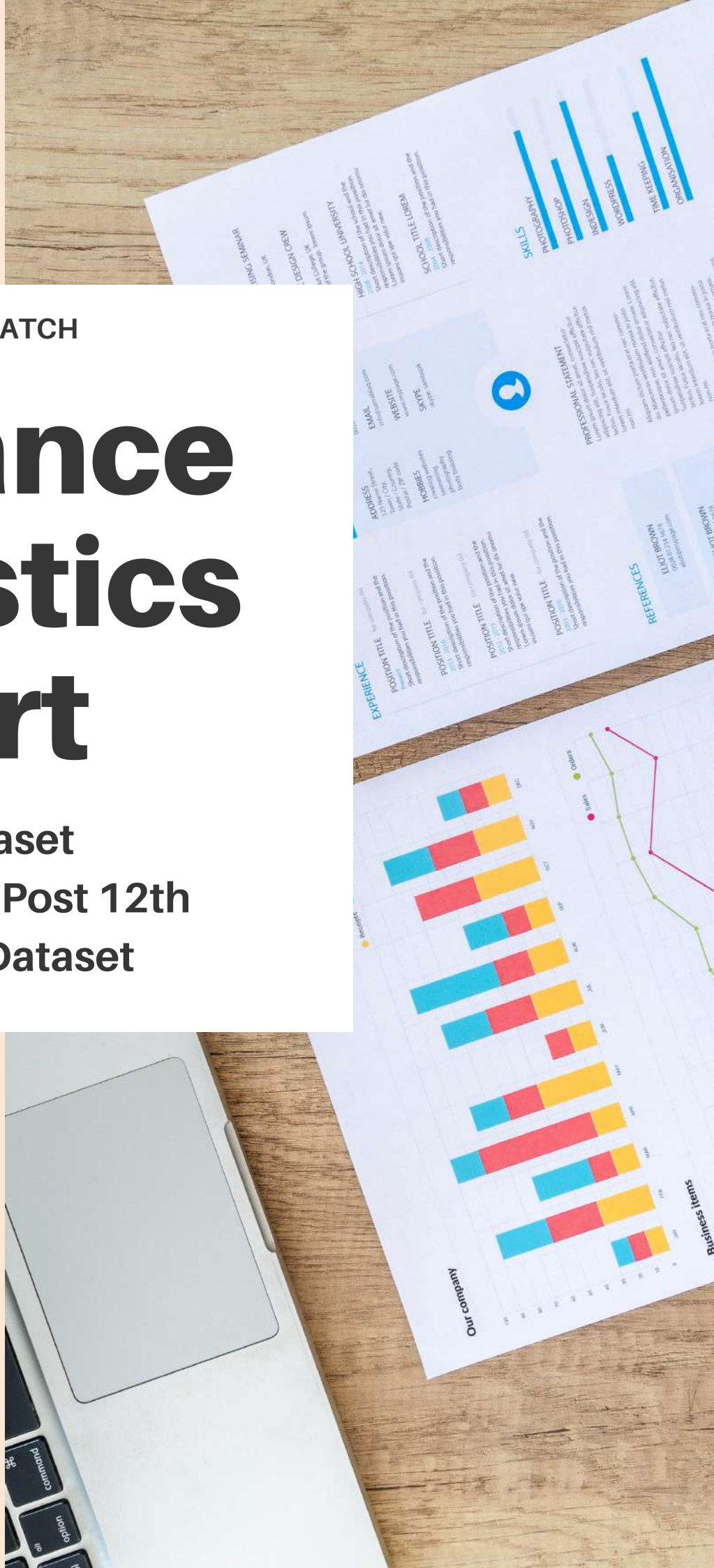
DSBA APRIL 2021 BATCH

Advance Statistics Report

- Salary Dataset
- Education Post 12th Standard Dataset



Prepared by Ruchita Abrange
July 2021

Table of Contents

Case Study "Salary"

Table of Contents

Case Study "Education post 12th Std"

About Case Study "Salary" for ANOVA

The salaries of 40 individuals 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.

Our Objective

A big project involves overseeing a lot of moving parts, oftentimes from different people. To have a successful rollout, project managers rely on a well-crafted project plan to ensure objectives are met on time and on budget. A project plan is a formal approved document that is used to define project goals, outline the project scope, monitor deliverables, and mitigate risks.

The objective of this dataset is that whether Jobs hypothesized to depend on educational qualification and occupation or not. To understand the dependency lets explore the dataset.

Data Description

1) Education :- Education taken by the candidates and it is categorized into 3 categories Doctorate, Bachelors and H.S Grade. It is in categorical data type.
2) Occupation :- Ocuupation is the occupation of them and it is categorized into 4 categories Adm-clerical, Sales, Prof-specialty and Exec-managerial. It is in categorical data type.
3) Salary :- The salary is the salary of them in integer data type.

Sample of the 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 SampleDataset has 3variables with 3 different varieties of education and 4 different occupation types.

Exploratory Data Analysis:

0 Education 40 non-null object
1 Occupation 40 non-null object
2 Salary 40 non-null int64

Let us check the types of variables in the data frame.

Dataset SampleDataset has 3 object datatype and one integer datatype. It doesn't have any null value.

Correlation Map

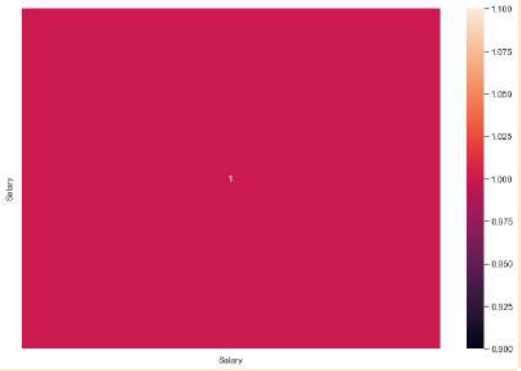


Fig. 1

Description

	Education	Occupation	Salary
count	40	40	40.00
unique	3	4	nan
top	Doctorate	Prof-specialty	nan
freq	16	13	nan
mean	NaN	NaN	162186.88
std	NaN	NaN	64860.41
min	NaN	NaN	50103.00
25%	NaN	NaN	99897.50
50%	NaN	NaN	169100.00
75%	NaN	NaN	214440.75
max	NaN	NaN	260151.00

Q1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

- H0: The Salary is depended on educational qualification.
- H1: The Salary is not depended on educational qualification.

Q1.2 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.

2. Perform the analysis of variances test on the dataset

	df	sum_sq	mean_sq	F	PR(>F)
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

3. Infere the test result

Since the p value is less than the significance level, we can reject the null hupothesis and state that Salary is not depended on educational qualification

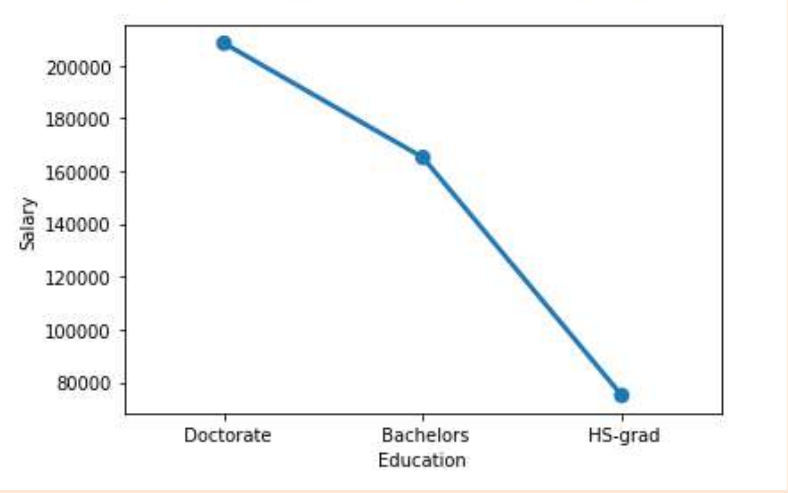


Fig. 2

```
Education     5.134773e+10
Residual      1.658718e+09
Name: mean_sq, dtype: float64
```

Q1.3 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.

Postulate the Null and Alternate Hypothesis for Salary to Occupation.

H_0: The Salary is depended on Occupation.

H_1: The Salary is not depended on Occupation.

	df	sum_sq	mean_sq	F	PR(>F)
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 the significance level, we can accept the null hupothesis and state that Salary is depended on Occupation

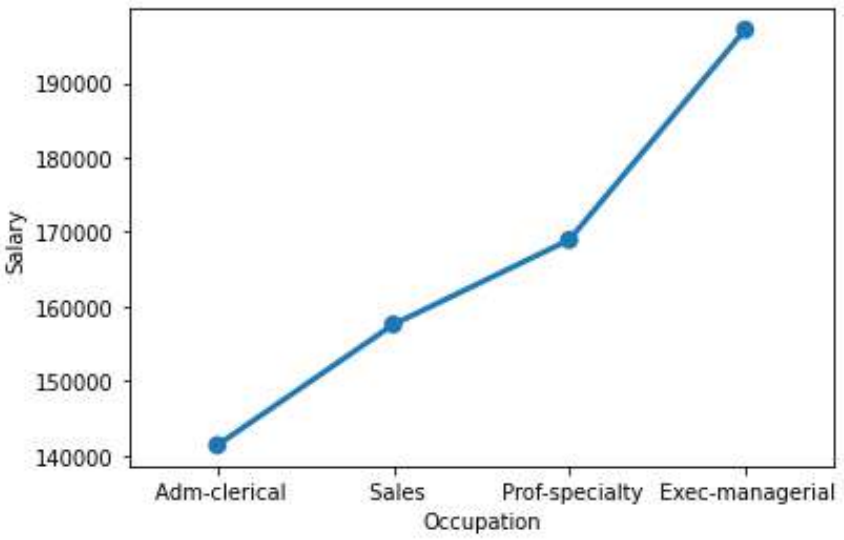


Fig. 3

Q1.4 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.

```
Occupation    3.752928e+09
Residual      4.244701e+09
Name: mean_sq, dtype: float64
```

Since mean_sq difference is more between Educational qualification class this means that Education class means are significantly different.

Q1.5) 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.

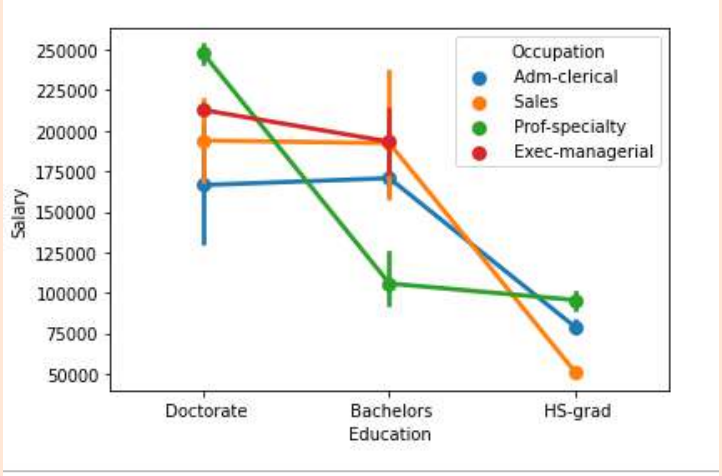
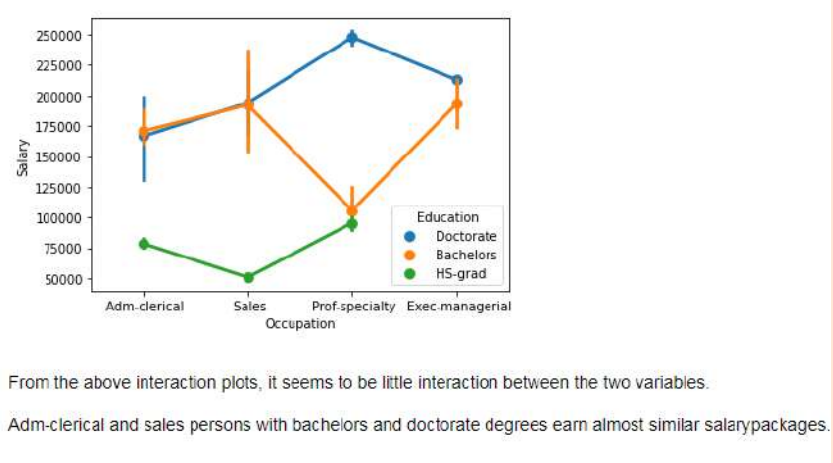


Fig. 4



From the above interaction plots, it seems to be little interaction between the two variables.

Adm-clerical and sales persons with bachelors and doctorate degrees earn almost similar salarypackages.

Fig. 5

- From the above interaction plots, it seems to be little interaction between the two variables.
- Adm-clerical and sales persons with bachelors and doctorate degrees earn almost similar salarypackages.

Q1.6 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?

H_0: The Salary is depended both on Education and Occupation.

H_1: The Salary is not depended both on Education and Occupation.

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

Since the p value is lesser than the significance level, we can accept the null hupothesis and state that Salary is depended both with respect to Education and Occupation

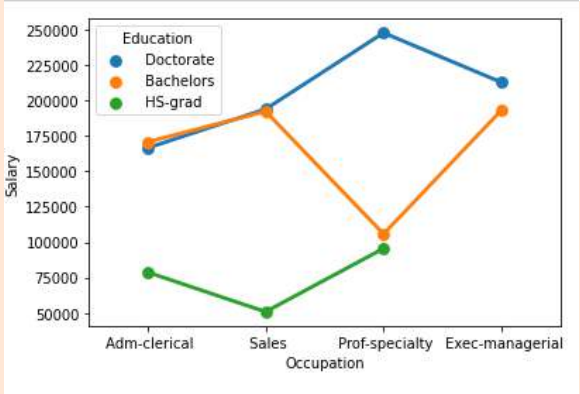


Fig. 6

Q1.7 Explain the business implications of performing ANOVA for this particular case study.

- By performing the ANOVA test for Salary case study we can say that Salary is not depended on educational qualification but infact it is little dependent on Occupation.
- But when considered both the class it says that Salary is moderately depended both with respect to Education and Occupation

About Case Study "Education post 12th Std" for EDA & PCA

The dataset Education - Post 12th Standard.csv contains information on various colleges.

Our Objective

A big project involves overseeing a lot of moving parts, oftentimes from different people. To have a successful rollout, project managers rely on a well-crafted project plan to ensure objectives are met on time and on budget. A project plan is a formal approved document that is used to define project goals, outline the project scope, monitor deliverables, and mitigate risks.

The objective of this dataset is that We are expected to do a Principal Component Analysis for this case study.

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

We start with loading the dataset, checking its shape and data types of variable . shape tell us how many rows and columns we have in the data and data type tell us whether the variable is object,integer or float value..

Sample of the dataset:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

Table 2.

```
RANGEINDEX: 777 ENTRIES, 0 TO 776
DATA COLUMNS (TOTAL 18 COLUMNS):
 #   COLUMN       NON-NULL COUNT  DTYPE
---  ------       --------------  -----
 0   NAMES        777 NON-NULL    OBJECT
 1   APPS         777 NON-NULL    INT64
 2   ACCEPT       777 NON-NULL    INT64
 3   ENROLL       777 NON-NULL    INT64
 4   TOP10PERC    777 NON-NULL    INT64
 5   TOP25PERC    777 NON-NULL    INT64
 6   F.UNDERGRAD  777 NON-NULL    INT64
 7   P.UNDERGRAD  777 NON-NULL    INT64
 8   OUTSTATE     777 NON-NULL    INT64
 9   ROOM.BOARD   777 NON-NULL    INT64
 10  BOOKS        777 NON-NULL    INT64
 11  PERSONAL     777 NON-NULL    INT64
 12  PHD          777 NON-NULL    INT64
 13  TERMINAL     777 NON-NULL    INT64
 14  S.F.RATIO    777 NON-NULL    FLOAT64
 15  PERC.ALUMNI  777 NON-NULL    INT64
 16  EXPEND       777 NON-NULL    INT64
 17  GRAD.RATE    777 NON-NULL    INT64

DTYPES: FLOAT64(1), INT64(16), OBJECT(1)
MEMORY USAGE: 109.4+ KB
```

The dataset Education - Post 12th Standard.csv contains non null values in dataset and 1 object type, 16 integer data type and 1 float datatype.

The dataset provided is for the student who enrol in the university / college after 12th std.

Column names Name, Accept, Enroll will be more meaningful if associated with S.F.Ratio, Grad.Rate. No duplicate records

Description

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	pe
count	777.00	777.00	777.00	777.00	777.00	777.00	777.00	777.00	777.00	777.00	777.00	777.00	777.00	777.00	
mean	3001.64	2018.80	779.97	27.56	55.80	3699.91	855.30	10440.67	4357.53	549.38	1340.64	72.66	79.70	14.09	
std	3870.20	2451.11	929.18	17.64	19.80	4850.42	1522.43	4023.02	1096.70	165.11	677.07	16.33	14.72	3.96	
min	81.00	72.00	35.00	1.00	9.00	139.00	1.00	2340.00	1780.00	96.00	250.00	8.00	24.00	2.50	
25%	776.00	604.00	242.00	15.00	41.00	992.00	95.00	7320.00	3597.00	470.00	850.00	62.00	71.00	11.50	
50%	1558.00	1110.00	434.00	23.00	54.00	1707.00	353.00	9990.00	4200.00	500.00	1200.00	75.00	82.00	13.60	
75%	3624.00	2424.00	902.00	35.00	69.00	4005.00	967.00	12925.00	5050.00	600.00	1700.00	85.00	92.00	16.50	
max	48094.00	26330.00	6392.00	96.00	100.00	31643.00	21836.00	21700.00	8124.00	2340.00	6800.00	103.00	100.00	39.80	

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

The main purpose of univariate analysis is to describe the data, summarize and finds pattern, it doesn't deal with causes and relationships unlike regression.

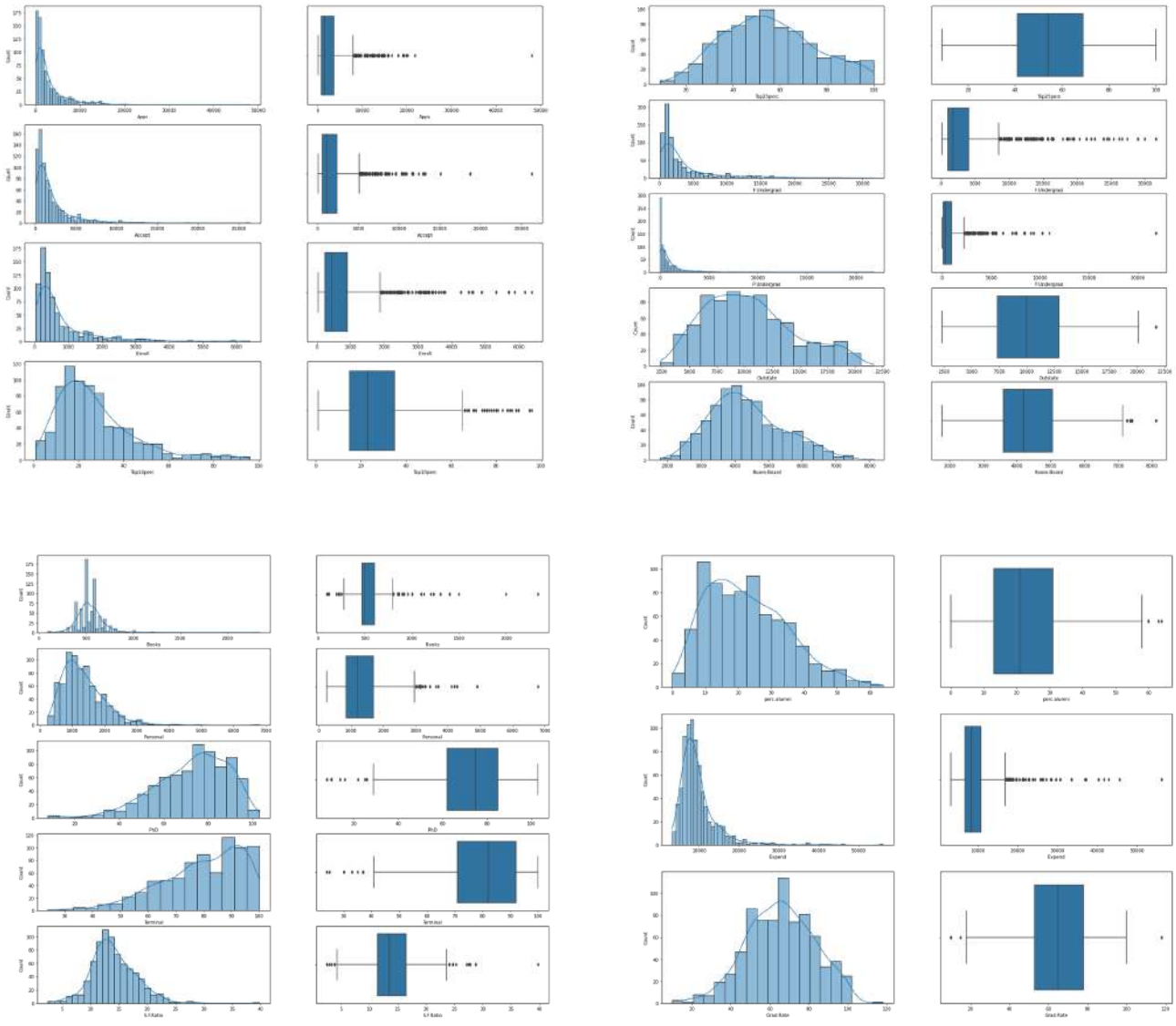


Fig. 7

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

Skweness of Data

1) Skweness = 0 -------- Normally Distributed

2) Skweness > 0 ---------Left Skewed

3) Skweness < 0 ---------Right Skewed.

From above we can say that Top25perc and Grad.Rate is almost normally distributed.

Multivariate Analysis

	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	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
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	0.337583	0.176229	-0.159990	0.124717	0.067313
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	0.308274	0.237271	-0.180794	0.064169	-0.022341
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	0.491135	-0.384875	0.455485	0.660913	0.494989
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	0.524749	-0.294629	0.417884	0.527447	0.477281
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	0.300019	0.279703	-0.229462	0.018652	-0.078773
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	0.141904	0.232531	-0.280792	-0.083568	-0.257001
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	0.407983	-0.554821	0.566262	0.672779	0.571290
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	0.374540	-0.362628	0.272363	0.501739	0.424942
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	0.099955	-0.031929	-0.040208	0.112409	0.001061
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	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
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	0.849587	-0.130530	0.249009	0.432762	0.305038
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	1.000000	-0.160104	0.267130	0.438799	0.289527
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	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417884	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.267130	-0.402929	1.000000	0.417712	0.490898
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	0.438799	-0.583832	0.417712	1.000000	0.390343
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	0.289527	-0.306710	0.490898	0.390343	1.000000

Table. 2

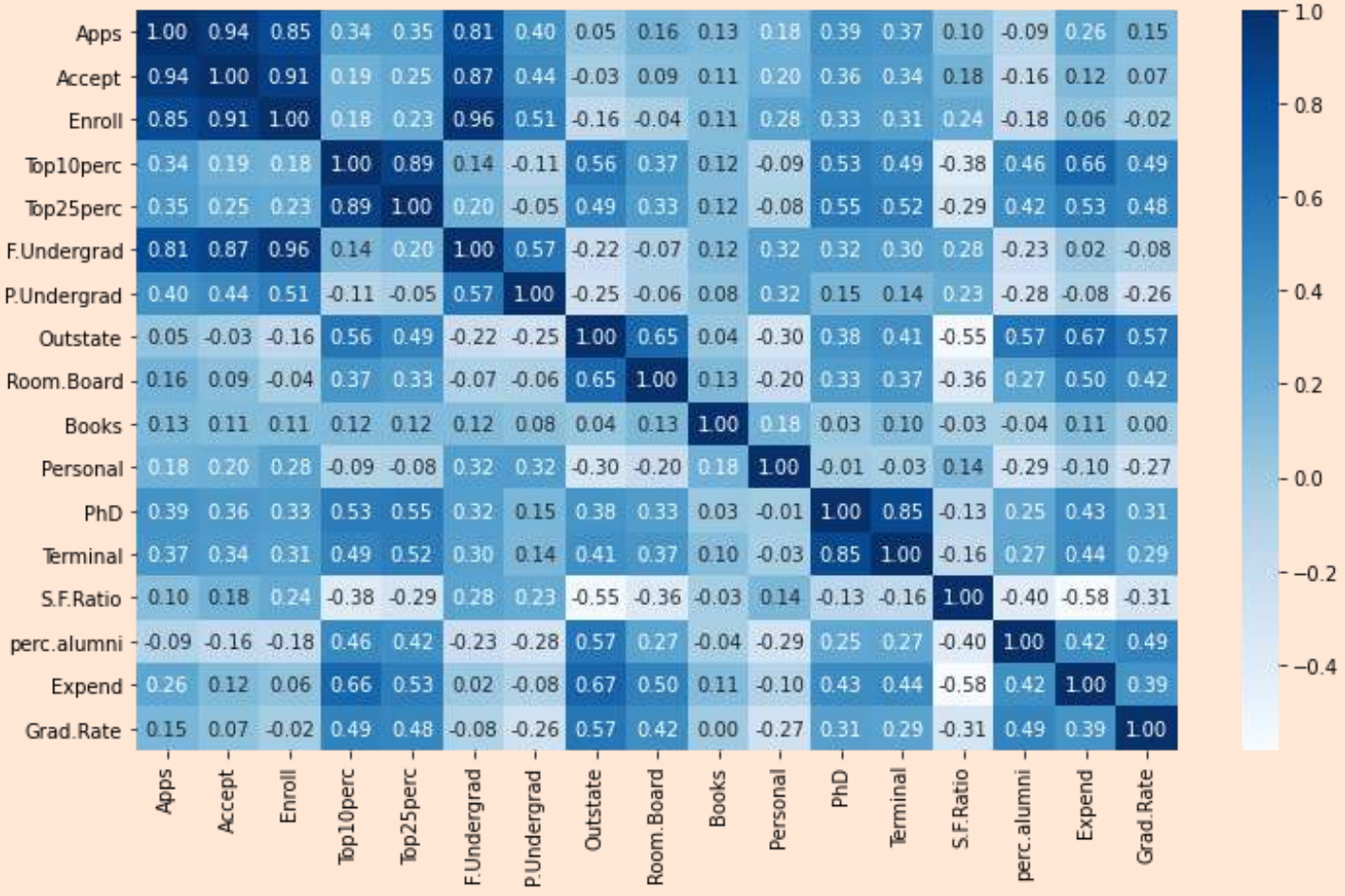


Fig. 8

A big project involves overseeing a lot of moving parts, oftentimes from different people. To have a successful rollout, project managers rely on a well-crafted project plan to ensure objectives are met on time and on budget. A project plan s a formal approved document which is used to define project goals, outline the project scope, monitor deliverables, and mitigate risks.

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

Yes it is necessary to do scaling for PCA in this case.

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set many variables are having values in thousands and in other just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale.

StandardScaler normalizes the data using the formula (x-mean)/standard deviation.

We will be doing this only for the numerical variables.*Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set Income is having values in thousands and age in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.**

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale.

StandardScaler normalizes the data using the formula (x-mean)/standard deviation.

We will be doing this only for the numerical variables.

Since scalling is done only on numerical values we will remove Names Column

We apply z score

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
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	-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
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
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
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

Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
-0.115729	1.013776	-0.867574	-0.501910	-0.318252
-3.378176	-0.477704	-0.544572	0.166110	-0.551262
-0.931341	-0.300749	0.585935	-0.177290	-0.667767
1.175657	-1.615274	1.151188	1.792851	-0.376504
-0.523535	-0.553542	-1.675079	0.241803	-2.939613

Tab. 3

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

Correlation is a scaled version of covariance; note that the two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be positively correlated; when the sign is negative, the variables are said to be negatively correlated; and when the sign is 0, the variables are said to be uncorrelated.

In a simple sense correlation, measures both the strength and direction of the linear relationship between two variables Covariance is a measure used to determine how much two variables change in tandem. It indicates the direction of the linear relationship between 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	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.095756	-0.090342	0.259927	0.146944
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	0.338018	0.176456	-0.160196	0.124878	0.067399
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	0.308671	0.237577	-0.181027	0.064252	-0.022370
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	0.491768	-0.385370	0.456072	0.661765	0.495627
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	0.525425	-0.295009	0.418403	0.528127	0.477896
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	0.300406	0.280064	-0.229758	0.018676	-0.078875
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	0.142086	0.232830	-0.281154	-0.083676	-0.257332
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	0.408509	-0.555536	0.566992	0.673646	0.572026
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	0.375022	-0.363095	0.272714	0.502386	0.425489
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	0.100084	-0.031970	-0.040260	0.112554	0.001062
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	-0.030653	0.136521	-0.286337	-0.098018	-0.269691
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	0.850682	-0.130698	0.249330	0.433319	0.305431
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	1.001289	-0.160310	0.267475	0.439365	0.289900
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	-0.160310	1.001289	-0.403448	-0.584584	-0.307106
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	0.267475	-0.403448	1.001289	0.418250	0.491530
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	0.439365	-0.584584	0.418250	1.001289	0.390846
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	0.289900	-0.307106	0.491530	0.390846	1.001289

```
cov_matrix= np.cov(dff1_num_scaled.T)
cov_matrix
```

```
array([[ 1.00128866,  0.94466636,  0.84791332,  0.33927032,  0.35209304,
         0.81554018,  0.3987775 ,  0.05022367,  0.16515151,  0.13272942,
         0.17896117,  0.39120081,  0.36996762,  0.09575627, -0.09034216,
         0.2599265 ,  0.14694372],
       [ 0.94466636,  1.00128866,  0.91281145,  0.19269493,  0.24779465,
         0.87534985,  0.44183938, -0.02578774,  0.09101577,  0.11367165,
         0.20124767,  0.35621633,  0.3380184 ,  0.17645611, -0.16019604,
         0.12487773,  0.06739929],
       [ 0.84791332,  0.91281145,  1.00128866,  0.18152715,  0.2270373 ,
         0.96588274,  0.51372977, -0.1556777 , -0.04028353,  0.11285614,
         0.28129148,  0.33189629,  0.30867133,  0.23757707, -0.18102711,
         0.06425192, -0.02236983],
       [ 0.33927032,  0.19269493,  0.18152715,  1.00128866,  0.89314445,
         0.1414708 , -0.10549205,  0.5630552 ,  0.37195909,  0.1190116 ,
        -0.09343665,  0.53251337,  0.49176793, -0.38537048,  0.45607223,
         0.6617651 ,  0.49562711],
       [ 0.35209304,  0.24779465,  0.2270373 ,  0.89314445,  1.00128866,
         0.19970167, -0.05364569,  0.49002449,  0.33191707,  0.115676  ,
        -0.08091441,  0.54656564,  0.52542506, -0.29500852,  0.41840277,
         0.52812713,  0.47789622],
       [ 0.81554018,  0.87534985,  0.96588274,  0.1414708 ,  0.19970167,
         1.00128866,  0.57124738, -0.21602002, -0.06897917,  0.11569867,
         0.31760831,  0.3187472 ,  0.30040557,  0.28006379, -0.22975792,
         0.01867565, -0.07887464],
       [ 0.3987775 ,  0.44183938,  0.51372977, -0.10549205, -0.05364569,
         0.57124738,  1.00128866, -0.25383901, -0.06140453,  0.08130416,
         0.32029384,  0.14930637,  0.14208644,  0.23283016, -0.28115421,
        -0.08367612, -0.25733218],
       [ 0.05022367, -0.02578774, -0.1556777 ,  0.5630552 ,  0.49002449,
        -0.21602002, -0.25383901,  1.00128866,  0.65509951,  0.03890494,
        -0.29947232,  0.38347594,  0.40850895, -0.55553625,  0.56699214,
         0.6736456 ,  0.57202613],
       [ 0.16515151,  0.09101577, -0.04028353,  0.37195909,  0.33191707,
        -0.06897917, -0.06140453,  0.65509951,  1.00128866,  0.12812787,
        -0.19968518,  0.32962651,  0.3750222 , -0.36309504,  0.27271444,
         0.50238599,  0.42548915],
       [ 0.13272942,  0.11367165,  0.11285614,  0.1190116 ,  0.115676  ,
         0.11569867,  0.08130416,  0.03890494,  0.12812787,  1.00128866,
         0.17952581,  0.0269404 ,  0.10008351, -0.03197042, -0.04025955,
         0.11255393,  0.00106226],
       [ 0.17896117,  0.20124767,  0.28129148, -0.09343665, -0.08091441,
         0.31760831,  0.32029384, -0.29947232, -0.19968518,  0.17952581,
```

```
dff1_num_scaled.corr()
```

	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

Tab. 4

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

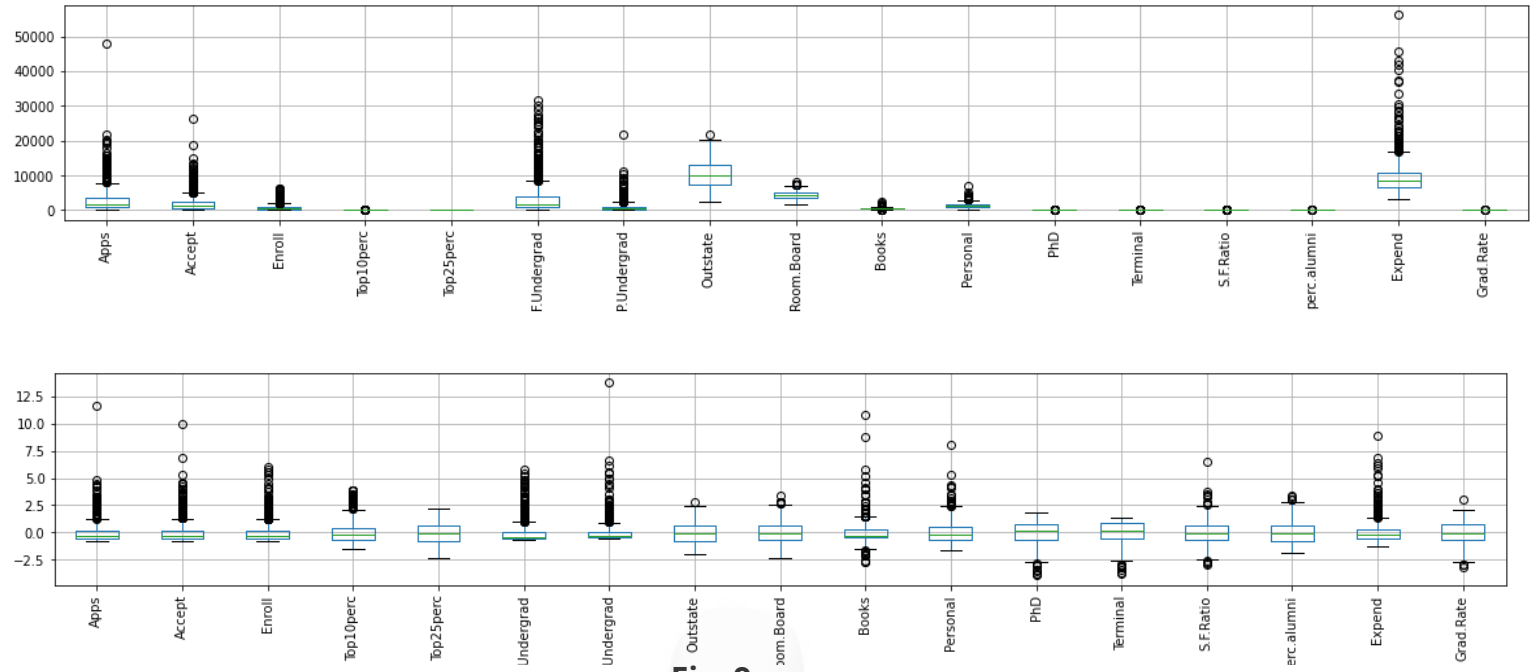


Fig. 9

We can see that there are many changes in outliers. Here can see that except for Top25perc column all other column has outliers

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

Eigenvalue and Eigenmatrix are mainly used to capture key information that stored in a large matrix.

1. It provides summary of large matrix.
2. Performing computation on large matrix is slow and require more memory and CPU, eigenvectors and eigenvalues can improve the efficiency in computationally intensive task by reducing dimensions after ensuring of the key information is maintained.

```
#Apply PCA taking all features
from sklearn.decomposition import PCA
pca = PCA(n_components=17, random_state=123)
pca_transformed = pca.fit_transform(df1_num_scaled)
```

```
#Extract eigen values
pca.explained_variance_
```

```
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])
```

```
#Extract eigen vectors
pca.components_
```

```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
         2.26743985e-01, -2.08064649e-01],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
         7.92734946e-02,  2.69129066e-01],
       [ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
        -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
```

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

To decide how many eigenvalues/eigenvectors to keep, we should clearly define the objective first for doing PCA in the first place. Are we doing it for reducing storage requirements, to reduce dimensionality for a classification algorithm, or for some other reason.

If we don't have any strict constraints, then we should plot the cumulative sum of eigenvalues.
If we divide each value by the total sum of eigenvalues prior to plotting, then your plot will show the fraction of total variance retained vs. a number of eigenvalues. The plot will then provide a good indication of when you hit the point of diminishing retur

```
]: # Apply PCA for the number of decided components to get the loadings and component output

   # Using scikit learn PCA here. It maps data to PCA dimensions in one shot
   from sklearn.decomposition import PCA
   # NOTE - we are generating only 8 PCA dimensions (dimensionality reduction from 18 to 7)
   pca = PCA(n_components=7, random_state=123)
   df_pca = pca.fit_transform(df1_num_scaled)
   df_pca.transpose() # Component output
```

```
]: array([[-1.59285540e+00, -2.19240180e+00, -1.43096371e+00, ...,
          -7.32560596e-01,  7.91932735e+00, -4.69508066e-01],
         [ 7.67333510e-01, -5.78829984e-01, -1.09281889e+00, ...,
          -7.72352397e-02, -2.06832886e+00,  3.66660943e-01],
         [-1.01073537e-01,  2.27879812e+00, -4.38092811e-01, ...,
          -4.05641899e-04,  2.07356368e+00, -1.32891515e+00],
         ...,
         [-7.43975398e-01,  1.05999660e+00, -3.69613274e-01, ...,
          -5.16021118e-01, -9.47754745e-01, -1.13217594e+00],
         [-2.98306081e-01, -1.77137309e-01, -9.60591689e-01, ...,
           4.68014248e-01, -2.06993738e+00,  8.39893087e-01],
         [ 6.38443468e-01,  2.36753302e-01, -2.48276091e-01, ...,
          -1.31749158e+00,  8.33276555e-02,  1.30731260e+00]])
```

```
]: #Check the explained variance for each PC
   #Note: Explained variance = (eigen value of each PC)/(sum of eigen values of all PCs)
   pca.explained_variance_ratio_
```

```
]: array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
         0.04984701, 0.03558871])
```

```
]: df_extracted_loadings = pd.DataFrame(pca.components_.T,
                                       columns = ['PC1','PC2', 'PC3', 'PC4', 'PC5', 'PC6',
                                                  'PC7'],
                                       index = df1_num_scaled.columns)
```

```
df_extracted_loadings
```

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

Scree plot: A scree plot helps the analyst visualize the relative importance of the factors, a sharp drop in the plot signals that subsequent factors are ignorable

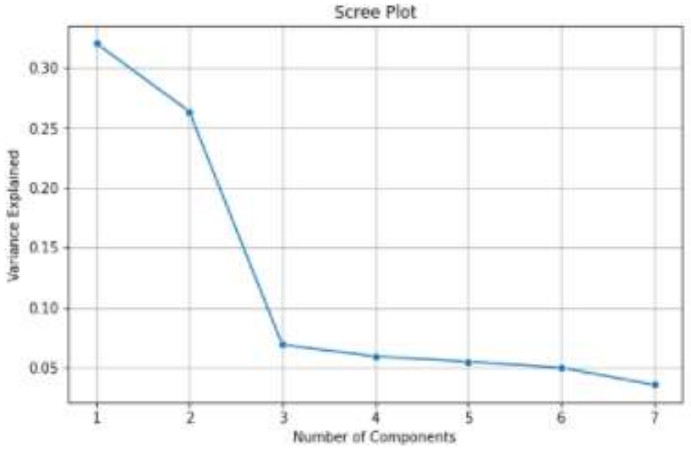


Fig. 10

Q2.7 Write down the explicit form of the first PC

	PC1	PC2	PC3	PC4	PC5
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919
Books	0.064758	0.056342	0.677412	0.087089	-0.127289
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268

Tab. 5

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?

We should define the objective first for doing PCA in the first place. Are we doing it for reducing storage requirements, to reduce dimensionality for a classification algorithm, or for some other reason.
If we don't have any strict constraints, then we should plot the cumulative sum of eigenvalues. If we divide each value by the total sum of eigenvalues prior to plotting, then your plot will show the fraction of total variance retained vs. number of eigenvalues. The plot will then provide a good indication of when you hit the point of diminishing returns.

```
#Check the cumlative explained variance ratio to find a cut off for selecting the number of PCs
np.cumsum(pca.explained_variance_ratio_)
```

```
array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
       0.81657854, 0.85216726])
```

```
var = np.cumsum(np.round(pca.explained_variance_ratio_,3))*100
var
```

```
array([32. , 58.3, 65.2, 71.1, 76.6, 81.6, 85.2])
```

The Cumulative % gives the percentage of variance accounted for by the n components. For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components. It helps in deciding the number of components by selecting the components which explained the high variance In the above array we see that the first feature explains 32% of the variance within our data set while the first two explain 58.3 and so on. If we employ 7 features we capture ~ 85.2% of the variance within the dataset.

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

1. PCA is a statistical technique and uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. PCA also is a tool to reduce multidimensional data to lower dimensions while retaining most of the information. Principal Component Analysis (PCA) is a well-established mathematical technique for reducing the dimensionality of data, while keeping as much variation as possible.
2. This PCA can only be done on continuous variables
3. There are about 18 variables in the dataset, by applying PCA we will reduce those to just 7 components which will capture 87.6 % variance in the dataset