18 OCTOBER 2020

PROJECT REPORT BY -
CHETAN DUDHANE

# ADVANCED STATISTICS

Registered Email : chetan.dudhane@gmail.com
Batch : PGP DSBA July B Group 2

# INTRODUCTION

This report consists of analysis of two problem statements -

- PROBLEM 1 - Effect of Compound on Relief in Hay Fever (ANOVA)
- PROBLEM 2 - Education post 12th (EDA and PCA)

Please find the Jupyter Code Notebook in the Google Drive link below. Analysis code is in Python. Datasets used are in the same directory. - https://bit.ly/3iSzp4g

# PROBLEM 1 - HAY FEVER

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomisation was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv

## 1.A EXPLORATORY ANALYSIS

| A | B | Volunteer | Relief |
|---|---|---|---|
| 1 | 1 | 1 | 2.4 |
| 1 | 1 | 2 | 2.7 |
| 1 | 1 | 3 | 2.3 |
| 1 | 1 | 4 | 2.5 |
| 1 | 2 | 1 | 4.6 |
| 1 | 2 | 2 | 4.2 |
| 1 | 2 | 3 | 4.9 |
| 1 | 2 | 4 | 4.7 |
| 1 | 3 | 1 | 4.8 |
| 1 | 3 | 2 | 4.5 |

Table. 1.1 - Fever Data (first 10 rows)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   A          36 non-null     int64
 1   B          36 non-null     int64
 2   Volunteer  36 non-null     int64
 3   Relief     36 non-null     float64
dtypes: float64(1), int64(3)
memory usage: 1.2 KB
```

Table 1.2 : Default Fever Data Info

## 1.B BASIC UNDERSTANDING OF DATA

1. Total No. Of Entries = **36**
2. Total No. Of Variables = 4
    - Data Type Float - 'RELIEF' in hours
    - Data Type Integer - 'A', 'B' and 'VOLUNTEER'
2. Study of **2** ingredients in the compound - **A and B**
    - Amount Levels in each ingredient = **3**
3. No. Of Treatments = 9
4. No. Of Volunteers to each Treatment = 4
5. We need to perform ANOVA on A and B, with and without interaction

- Hence, *we convert data type of variables 'A', 'B' and 'Volunteer' to* ***Categorical***.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   A          36 non-null     category
 1   B          36 non-null     category
 2   Volunteer  36 non-null     category
 3   Relief     36 non-null     float64
dtypes: category(3), float64(1)
memory usage: 924.0 bytes
```

Table 1.3 : Fever Data Info (After Converting to Categorical)

# 1.C DESCRIPTIVE ANALYSIS

| | A | B | Volunteer | Relief |
|---|---|---|---|---|
| **count** | 36.00 | 36.00 | 36.00 | 36.00 |
| **unique** | 3.00 | 3.00 | 4.00 | |
| **top** | 3.00 | 3.00 | 4.00 | |
| **freq** | 12.00 | 12.00 | 9.00 | |
| **mean** | | | | 7.18 |
| **std** | | | | 3.27 |
| **min** | | | | 2.30 |
| **0.25** | | | | 4.68 |
| **0.50** | | | | 6.00 |
| **0.75** | | | | 9.33 |
| **max** | | | | 13.50 |

Table 1.4 : Fever data - Statistical Description

1. There is NO Missing Data
2. There are NO Duplicate Rows
3. Due to the treatment, Relief varies from **2.3 hrs to 13.5 hrs**
4. All treatments together have a **Mean Relief = 7.18 hrs**

   with **Standard Deviation = 3.27 hrs**
5. Response Variable or **Dependent** Variable = **"Relief"**
6. Input Variables or **Independent** Variables = **"A", "B" and "Volunteer"**

**NOTE** : As given, We assume all ANOVA assumptions are satisfied

## [Q 1.1] State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually. [both statement and statistical form like Ho=mu, Ha>mu]

- **Treatment 1 - Effect of only Ingredient 'A' on 'Relief'**
  - NULL HYPOTHESIS : Means of RELIEF variable is SAME - due to different Levels of Ingredient A

$$H_O : \mu_{A1} = \mu_{A2} = \mu_{A3}$$

  - ALTERNATIVE HYPOTHESIS : At least One of the Means of RELIEF variable is not same to others - due to different levels of Ingredient A

$H_a$ : There is at least one pair of means, which is not same

- **Treatment 2 - Effect of only Ingredient 'B' on 'Relief'**
  - NULL HYPOTHESIS : Means of RELIEF variable is SAME - due to different Levels of Ingredient B

$$H_O : \mu_{B1} = \mu_{B2} = \mu_{B3}$$

  - ALTERNATIVE HYPOTHESIS : At least One of the Means of RELIEF variable is NOT SAME to others - due to different levels of Ingredient B

$H_a$ : There is at least one pair of means, which is not same

## [Q 1.2] Perform one-way ANOVA for variable 'A' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

**ONE WAY ANOVA - Effect of only Ingredient 'A' on 'Relief'**

**STEP 1 : STATE HYPOTHESIS**

- NULL HYPOTHESIS : Means of RELIEF variable is SAME - due to different Levels of Ingredient A

$$H_O : \mu_{A1} = \mu_{A2} = \mu_{A3}$$

- **ALTERNATIVE HYPOTHESIS : At least One of the Means of RELIEF variable is not same to others - due to different levels of Ingredient A**

  $H_a$ : There is at least one pair of means, which is not same

**STEP 2 : FIT THE FORMULA TO MODEL AND MAKE AN ANOVA TABLE**

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(A)** | 2.00 | 220.02 | 110.01 | 23.47 | 0.0000005 |
| **Residual** | 33.00 | 154.71 | 4.69 | | |

Table 1.5 : ANOVA TABLE - Effect of A on Relief

**STEP 3 : CONCLUDE AND INFER FROM THE ANOVA TABLE**

- As its not given, assuming industry standard of 95% confidence level. Hence,

  Level of Significance, $\alpha = 0.05$

- **p value** = 4.5782418430432463e-07 = **0.0000005**
- **Here, $\mathbf{p\ value} < \alpha$ (Level of Significance)**
- **We have enough evidence to reject the Null Hypothesis in favour of Alternative Hypothesis**
- **There is at least one pair of Means of RELIEF, which is NOT SAME.**
- **WE CONCLUDE THAT DIFFERENT LEVELS OF INGREDIENT A HAVE SIGNIFICANT IMPACT ON MEAN RELIEF.**

## [Q 1.3] Perform one-way ANOVA for variable 'B' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

**ONE WAY ANOVA - Effect of only Ingredient 'B' on 'Relief'**

**STEP 1 : STATE HYPOTHESIS**

- **NULL HYPOTHESIS : Means of RELIEF variable is SAME - due to different Levels of Ingredient B**

$$H_O : \mu_{B1} = \mu_{B2} = \mu_{B3}$$

- **ALTERNATIVE HYPOTHESIS : At least One of the Means of RELIEF variable is not same to others - due to different levels of Ingredient B**

$H_a$ **:** There is at least one pair of means, which is not same

**STEP 2 : FIT THE FORMULA TO MODEL AND MAKE AN ANOVA TABLE**

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(B)** | 2.00 | 123.66 | 61.83 | 8.13 | 0.0013498 |
| **Residual** | 33.00 | 251.07 | 7.61 | | |

Table 1.6 : ANOVA TABLE - Effect of B on Relief

**STEP 3 : CONCLUDE AND INFER FROM THE ANOVA TABLE**

- As its not given, assuming industry standard of 95% confidence level. Hence,

Level of Significance, $\alpha = 0.05$

- **p value = 0.0013498**
- **Here,** $\mathbf{p\ value} < \alpha$ **(Level of Significance)**
- **We have enough evidence to reject the Null Hypothesis in favour of Alternative Hypothesis**
- **There is at least one pair of Means of RELIEF, which is NOT SAME.**
- **WE CONCLUDE THAT DIFFERENT LEVELS OF INGREDIENT B HAVE SIGNIFICANT IMPACT ON MEAN RELIEF.**

## [Q 1.4] Analyse the effects of one variable on another with the help of an interaction plot.¶

### What is the interaction between the two treatments?

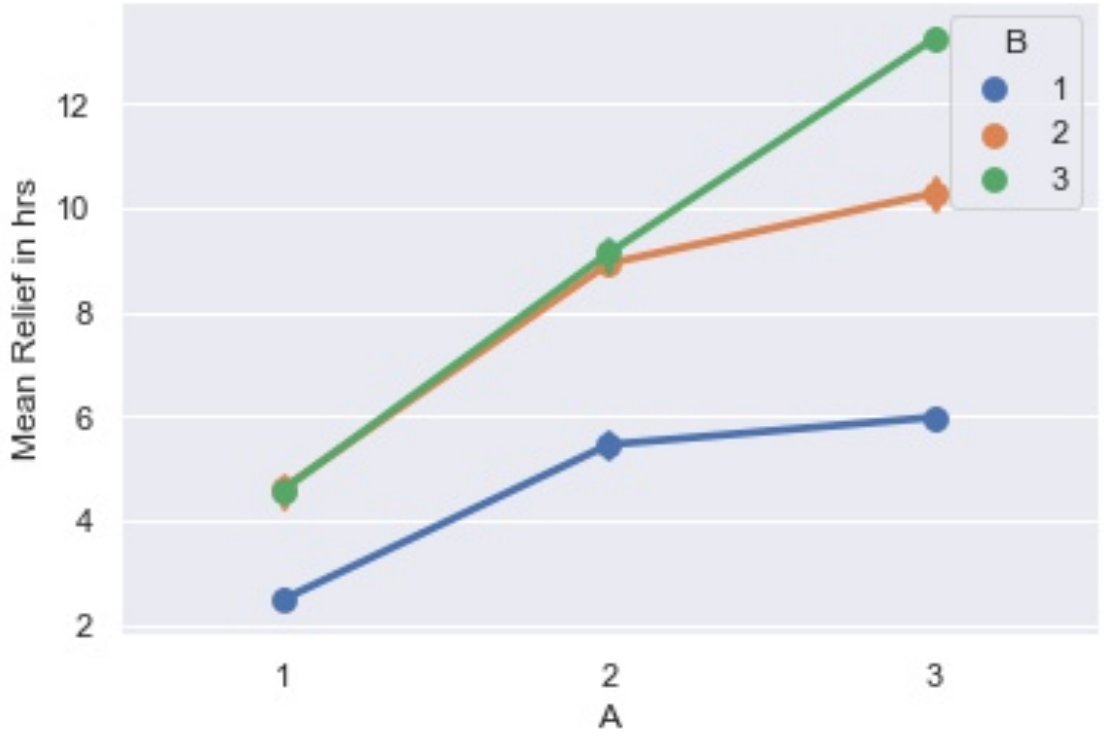


Fig 1.a : Combined Effect of Ingredients on Mean Relief

- There is a clear interaction between the 2 treatments (Effects of Ingredient A and B on Relief)
- Let different levels of A be known as A1, A2 and A3 and similar with B
- It is evident that treatment with **A1 & B1** has the **least** Relief, whereas the treatment with **A3 & B3** has the **maximum** effect with respect to Relief experienced by volunteers.
- Also, the Blue line above, it indicates B1 over 3 levels of A. We see that there is steep increase from A1 to A2 and not much difference later with A3
- B2 and B3 have exactly the same effect on Relief over A1 to A2.
- But from A2 to A3, B3 maintains the slope and rises higher but B2 slows down.
- **To summarise, as the amounts of Ingredients A and B in the Compound increase, the Relief also increases.**

## [Q 1.5] Perform a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’ along with their interaction 'A*B') with the variable 'Relief' and state your results.

**TWO WAY ANOVA - Effect Of Ingredients "A" And "B" And Its Interaction On "RELIEF"**

**STEP 1 : STATE THE HYPOTHESIS**

- **Treatment 1 - Effect of Ingredient 'A' on ‘Relief'**
  - NULL HYPOTHESIS : Means of RELIEF variable is SAME - due to different Levels of Ingredient A

    $$H_O : \mu_{A1} = \mu_{A2} = \mu_{A3}$$

  - ALTERNATIVE HYPOTHESIS : At least One of the Means of RELIEF variable is not same to others - due to different levels of Ingredient A

    $H_a$ : There is at least one pair of means, which is not same

- **Treatment 2 - Effect of Ingredient 'B' on ‘Relief'**
  - NULL HYPOTHESIS : Means of RELIEF variable is SAME - due to different Levels of Ingredient B

    $$H_O : \mu_{B1} = \mu_{B2} = \mu_{B3}$$

  - ALTERNATIVE HYPOTHESIS : At least One of the Means of RELIEF variable is NOT SAME to others - due to different levels of Ingredient B

    $H_a$ : There is at least one pair of means, which is not same

- **Effect of Interaction between Ingredients A and B on 'Relief'**
  - NULL HYPOTHESIS :

    $H_O$ : Interaction between Ingredients A and B DOES NOT significantly affect RELIEF variable

  - ALTERNATIVE HYPOTHESIS :

    $H_a$ : Interaction between Ingredients A and B significantly affects RELIEF variable

## STEP 2 : FIT THE FORMULA TO MODEL AND MAKE AN ANOVA TABLE

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(A)** | 2.00 | 220.02 | 110.01 | 1827.86 | 1.51404317443927E-29 |
| **C(B)** | 2.00 | 123.66 | 61.83 | 1027.33 | 3.34875124371216E-26 |
| **C(A):C(B)** | 4.00 | 29.43 | 7.36 | 122.23 | 6.9720832729235E-17 |
| **Residual** | 27.00 | 1.63 | 0.06 | | |

Table 1.6 : ANOVA TABLE - Effect of both A and B with Interaction on Relief

## STEP 3 : CONCLUDE AND INFER FROM THE ANOVA TABLE

- As its not given, assuming industry standard of 95% confidence level. Hence,

  Level of Significance, $\alpha = 0.05$

- **p-value of A = 1.51404317443927E-29. = 0.00000**

  p-value of A. <. $\alpha$

  We reject the Null Hypothesis of Treatment 1 - Effect of A on Relief

  **We conclude that different levels of Ingredient A have significant impact on Mean RELIEF**

- **p-value of B = 3.34875124371216E-26. = 0.00000**

  p-value of B. <. $\alpha$

  We reject the Null Hypothesis of Treatment 2 - Effect of B on Relief

  **We conclude that different levels of Ingredient B have significant impact on Mean RELIEF**

- **p-value of Interaction between A and B = 6.97208327292351E-17 = 0.00000**

  p-value of Interaction between A and B <. $\alpha$

  We have enough evidence to reject the Null Hypothesis of - Effect of Interaction between Ingredients A and B on 'Relief'

  **WE CONCLUDE THAT INTERACTION BETWEEN INGREDIENTS A AND B HAS SIGNIFICANT IMPACT ON MEAN 'RELIEF'**

## [Q 1.6] Mention the business implications of performing ANOVA for this particular case study.

- We performed One Way and Two Way ANOVA to determine the changes in the Response Variable 'RELIEF' caused by different amounts of Ingredients A and B in the compound.
- We understood that, individually, different amounts of ingredients A and B have a significant impact on 'RELIEF' experienced by the volunteers.
- Also, combined A & B has a significant effect on RELIEF. As we change the proportions of A & B in the compound, Relief also changes.
- **Generally speaking, we learn that Increasing amounts of A and B in the compound gives more and more Relief**
- Hence now, with these understandings -
    - Doctors can thus recommend treatments specifically dependent on patient's need for relief and her level of fever.
    - Like, a more critical patient can be prescribed the compound (A3 & B3), while someone just showing symptoms can be prescribed the compound of (A1 & B1)
    - Researchers can do away with the compounds (A1 - B3) and (A2 - B3), as they show the same effect on Relief as (A1 - B2) and (A2 - B2) respectively.
    - This study will help in pricing too. The compound (A3 - B3), which gives maximum Relief can be priced highest.
    - While, Compound (A2 - B2) can be mid-level priced.

# PROBLEM 2 - EDUCATION POST 12TH

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

## 2.A EXPLORATORY ANALYSIS

| Names | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad |
|---|---|---|---|---|---|---|---|
| **Abilene Christian University** | 660.0( | 1232.00 | 721.00 | 23.00 | 52 | **2885** | 537.00 |
| **Adelphi University** | 186.0( | 1924.00 | 512.00 | 16.00 | 29 | **2683** | 1227.00 |
| **Adrian College** | 428.0( | 1097.00 | 336.00 | 22.00 | 50 | **1036** | 99.00 |
| **Agnes Scott College** | 417.00 | 349.00 | 137.00 | 60.00 | 89 | **510** | 63.00 |
| **Alaska Pacific University** | 193.00 | 146.00 | 55.00 | 16.00 | 44 | **249.00** | 869.00 |

| Outstate | Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|---|---|---|---|---|---|
| **7440** | 3300.00 | 450.00 | 2200.00 | 70.00 | 78 | **18.1** | 12.00 | 7041.00 | 60.00 |
| **12280** | 6450.00 | 750.00 | 1500.00 | 29.00 | 30 | **12.2** | 16.00 | 10527.00 | 56.00 |
| **11250** | 3750.00 | 400.00 | 1165.00 | 53.00 | 66 | **12.9** | 30.00 | 8735.00 | 54.00 |
| **12960** | 5450.00 | 450.00 | 875.00 | 92.00 | 97 | **7.7** | 37.00 | 19016.00 | 59.00 |
| **7560.00** | 4120.00 | 800.00 | 1500.00 | 76.00 | 72 | **11.90** | 2.00 | **10922** | 15.00 |

Table 2.1 : First 5 rows of Education Data (Split Table)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Names        777 non-null    object
 1   Apps         777 non-null    int64
 2   Accept       777 non-null    int64
 3   Enroll       777 non-null    int64
 4   Top10perc    777 non-null    int64
 5   Top25perc    777 non-null    int64
 6   F.Undergrad  777 non-null    int64
 7   P.Undergrad  777 non-null    int64
 8   Outstate     777 non-null    int64
 9   Room.Board   777 non-null    int64
 10  Books        777 non-null    int64
 11  Personal     777 non-null    int64
 12  PhD          777 non-null    int64
 13  Terminal     777 non-null    int64
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64
 16  Expend       777 non-null    int64
 17  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

Table 2.2 : Education Data - Summary Info

## 2.B BASIC UNDERSTANDING OF DATA

1. Total No. of Entries (No. of Colleges in data) = **777**

2. Total No. of Variables (No. of Columns) = **18**

   a. No. of Continuous Variables = **17 ['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']**

   - Integer data type variables = 16
   - Float data type variables = 1. **['S.FRatio]**

   b. No. of Categorical Variables = **1 ['Names]**

3. There is NO Missing Data

4. There are NO Duplicate records

## 2.C DESCRIPTIVE ANALYSIS OF DATA

| | count | mean | std | min | 0.25 | 0.50 | 0.75 | max |
|---|---|---|---|---|---|---|---|---|
| **Apps** | 777 | 3001.64 | 3870.20 | 81 | 776 | 1558 | 3624 | 48094 |
| **Accept** | 777 | 2018.80 | 2451.11 | 72 | 604 | 1110 | 2424 | 26330 |
| **Enroll** | 777 | 779.97 | 929.18 | 35 | 242 | 434 | 902 | 6392 |
| **Top10perc** | 777 | 27.56 | 17.64 | 1 | 15 | 23 | 35 | 96 |
| **Top25perc** | 777 | 55.80 | 19.80 | 9 | 41 | 54 | 69 | 100 |
| **F.Undergrad** | 777 | 3699.91 | 4850.42 | 139 | 992 | 1707 | 4005 | 31643 |
| **P.Undergrad** | 777 | 855.30 | 1522.43 | 1 | 95 | 353 | 967 | 21836 |
| **Outstate** | 777 | 10440.67 | 4023.02 | 2340 | 7320 | 9990 | 12925 | 21700 |
| **Room.Board** | 777 | 4357.53 | 1096.70 | 1780 | 3597 | 4200 | 5050 | 8124 |
| **Books** | 777 | 549.38 | 165.11 | 96 | 470 | 500 | 600 | 2340 |
| **Personal** | 777 | 1340.64 | 677.07 | 250 | 850 | 1200 | 1700 | 6800 |
| **PhD** | 777 | 72.66 | 16.33 | 8 | 62 | 75 | 85 | 103 |
| **Terminal** | 777 | 79.70 | 14.72 | 24 | 71 | 82 | 92 | 100 |
| **S.F.Ratio** | 777 | 14.09 | 3.96 | 3 | 12 | 14 | 17 | 40 |
| **perc.alumni** | 777 | 22.74 | 12.39 | 0 | 13 | 21 | 31 | 64 |
| **Expend** | 777 | 9660.17 | 5221.77 | 3186 | 6751 | 8377 | 10830 | 56233 |
| **Grad.Rate** | 777 | 65.46 | 17.18 | 10 | 53 | 65 | 78 | 118 |

Table 2.3 : Education Data - Statistical Description

## [Q 2.1] Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented

### UNIVARIATE ANALYSIS

The following Analysis is done on raw data ***Before Treating Outliers.***

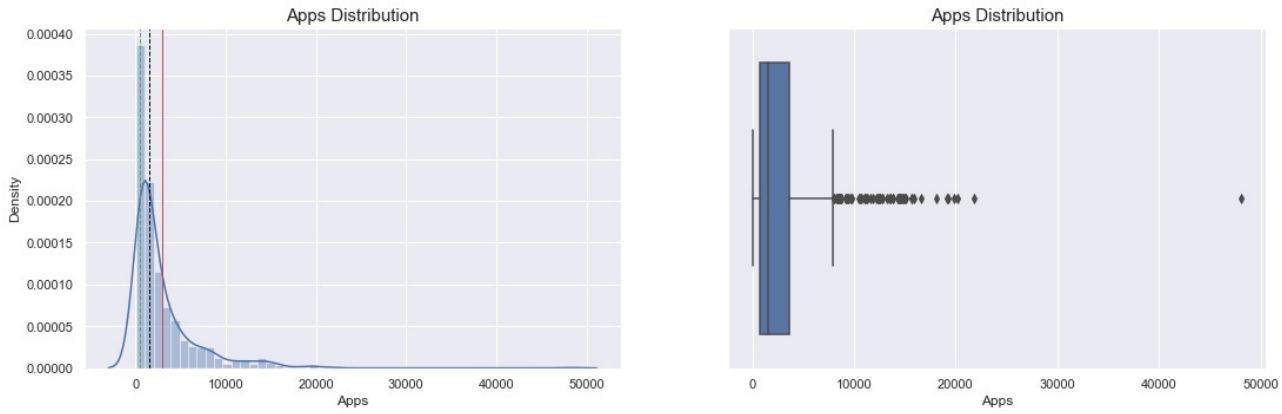


Fig 2.a : Dist and Box Plot - Apps

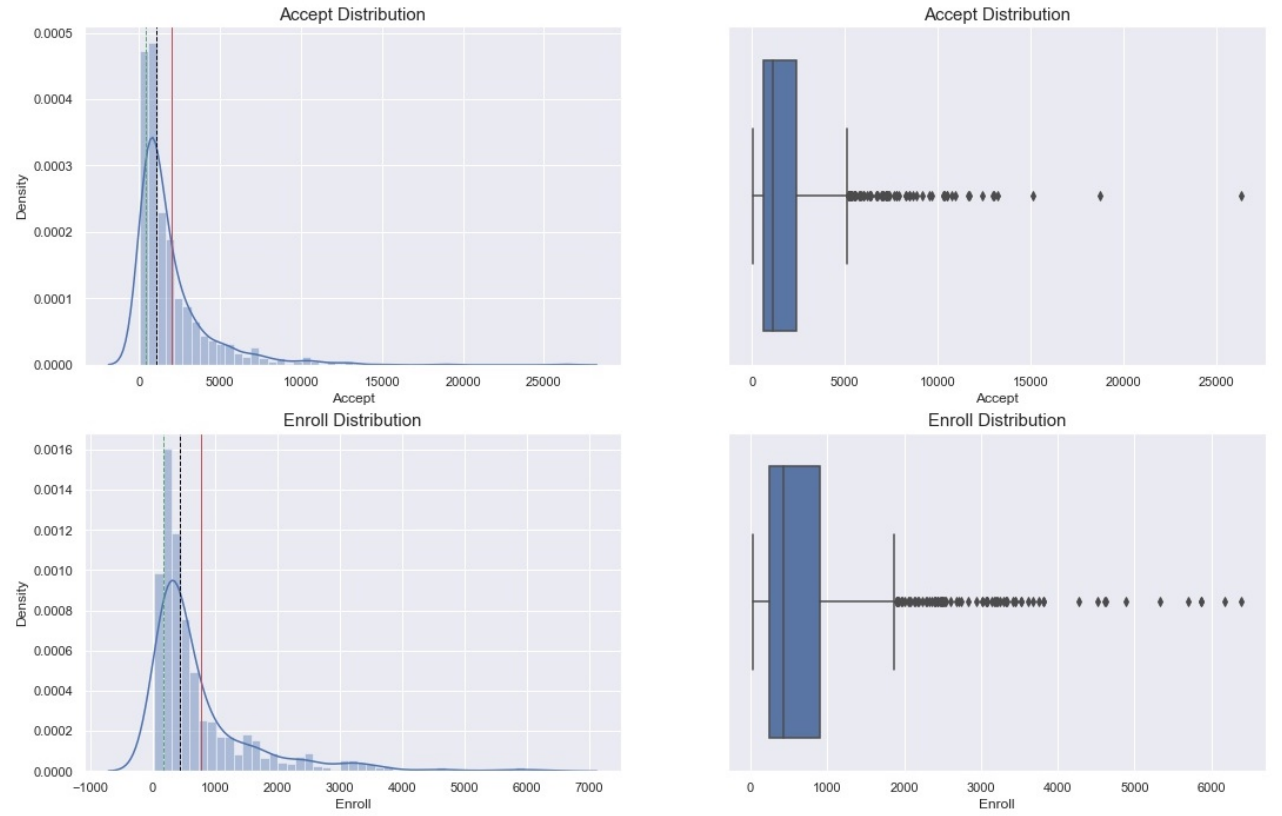


Fig 2.b : Dist and Box Plot - Accept and Enroll

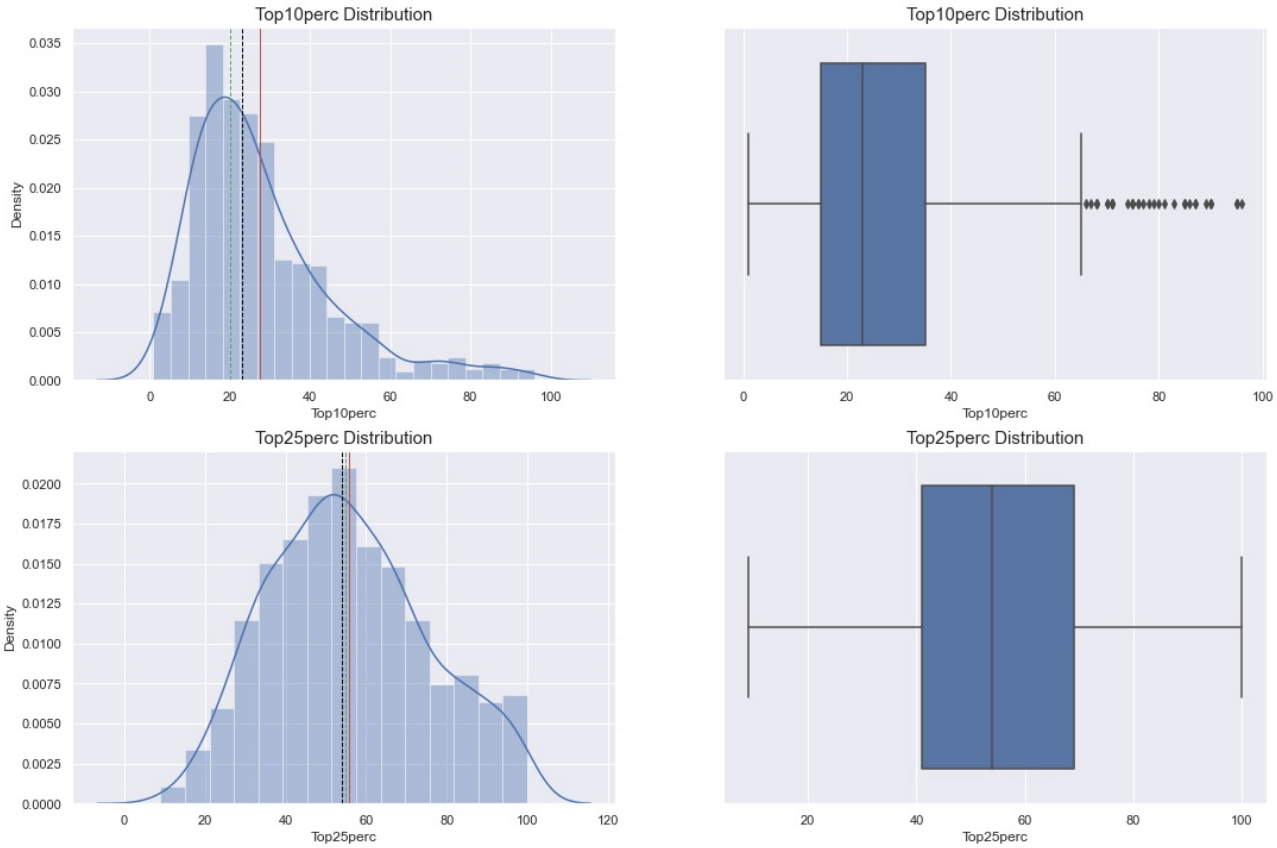


Fig 2.c : Dist and Box Plot - Top10perc & Top25perc

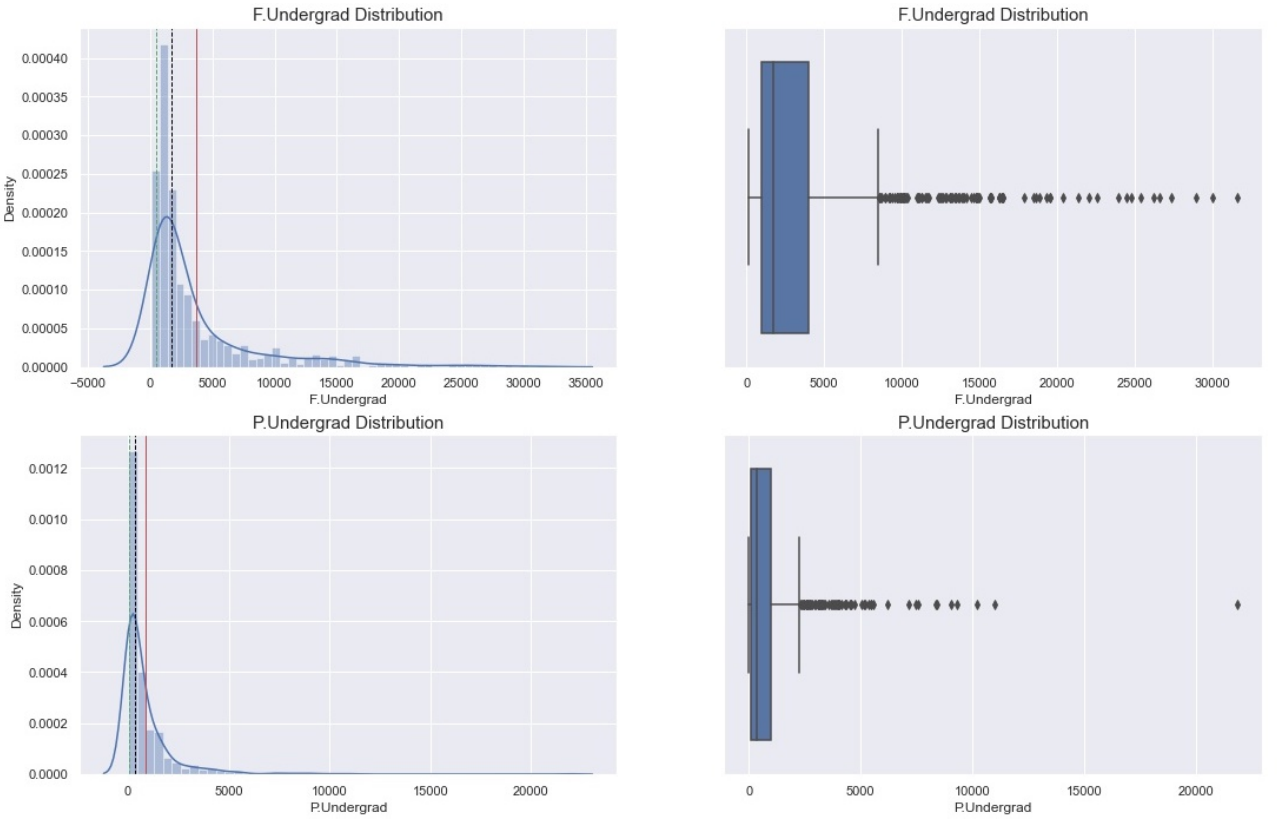


Fig 2.d.1 : Dist and Box Plot - F.Undergrad & P.Undergrad

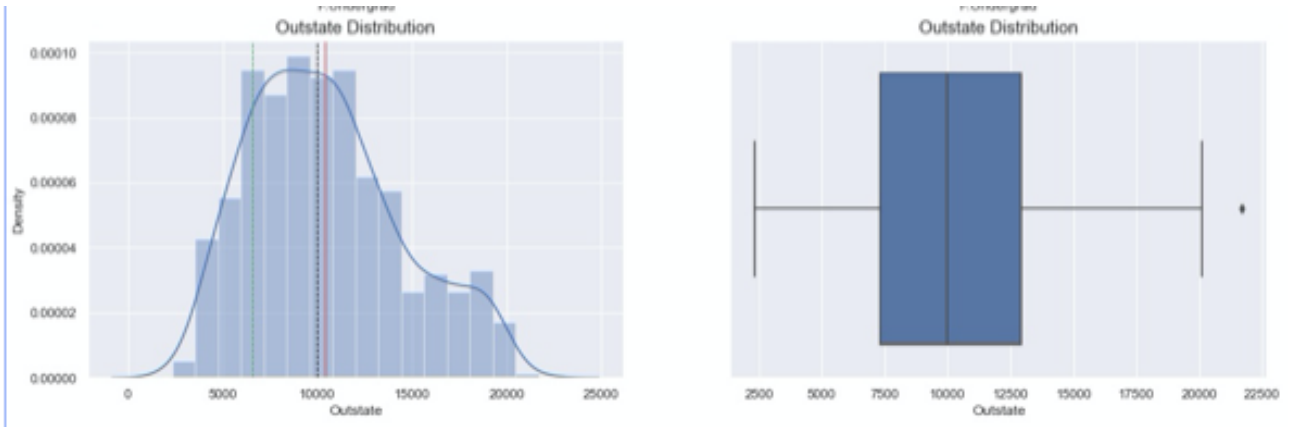


Fig 2.d.2 : Dist and Box Plot - Outstate

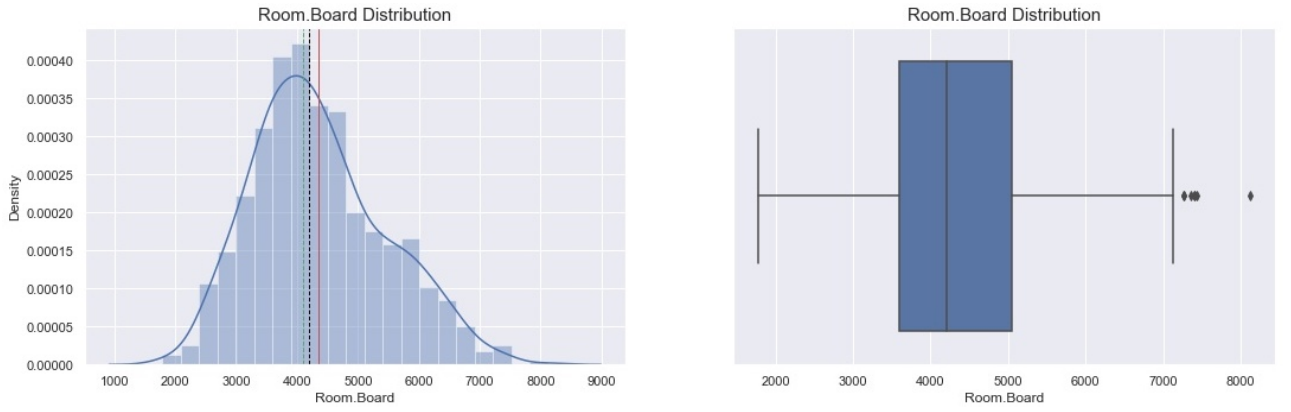


Fig 2.e : Dist and Box Plot - Expense Room board

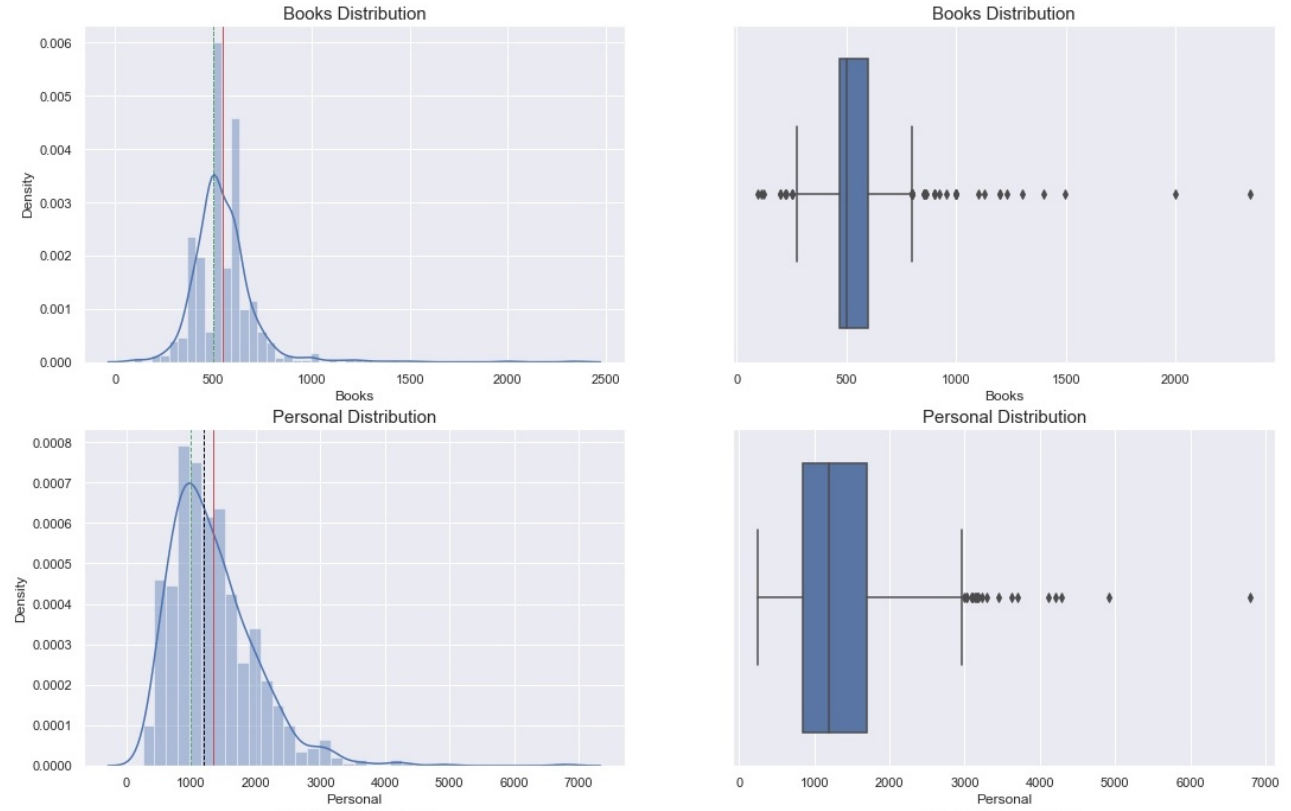


Fig 2.f : Dist and Box Plot - Expense Books and Personal

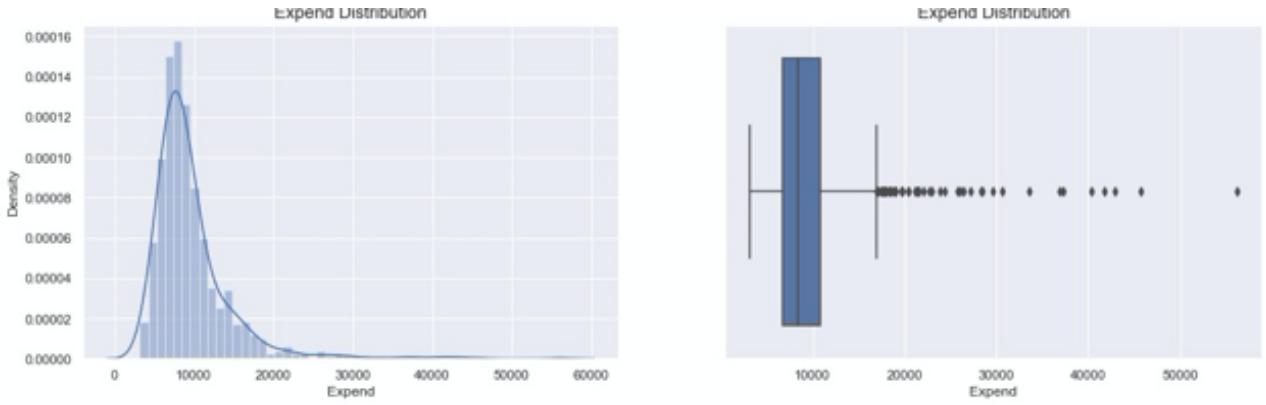


Fig 2.g : Dist and Box Plot - Instructional Expense

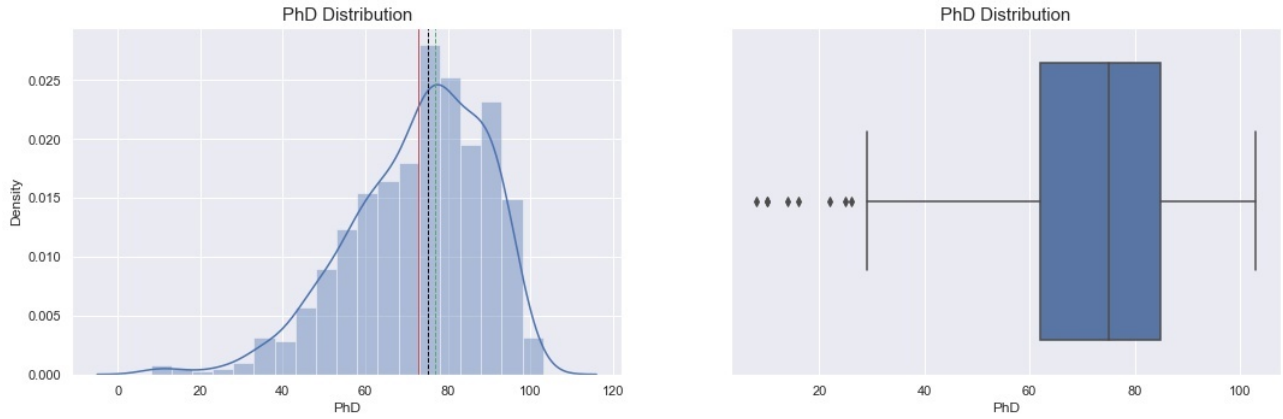


Fig 2.h.1 : Dist and Box Plot - Percent of Faculty with

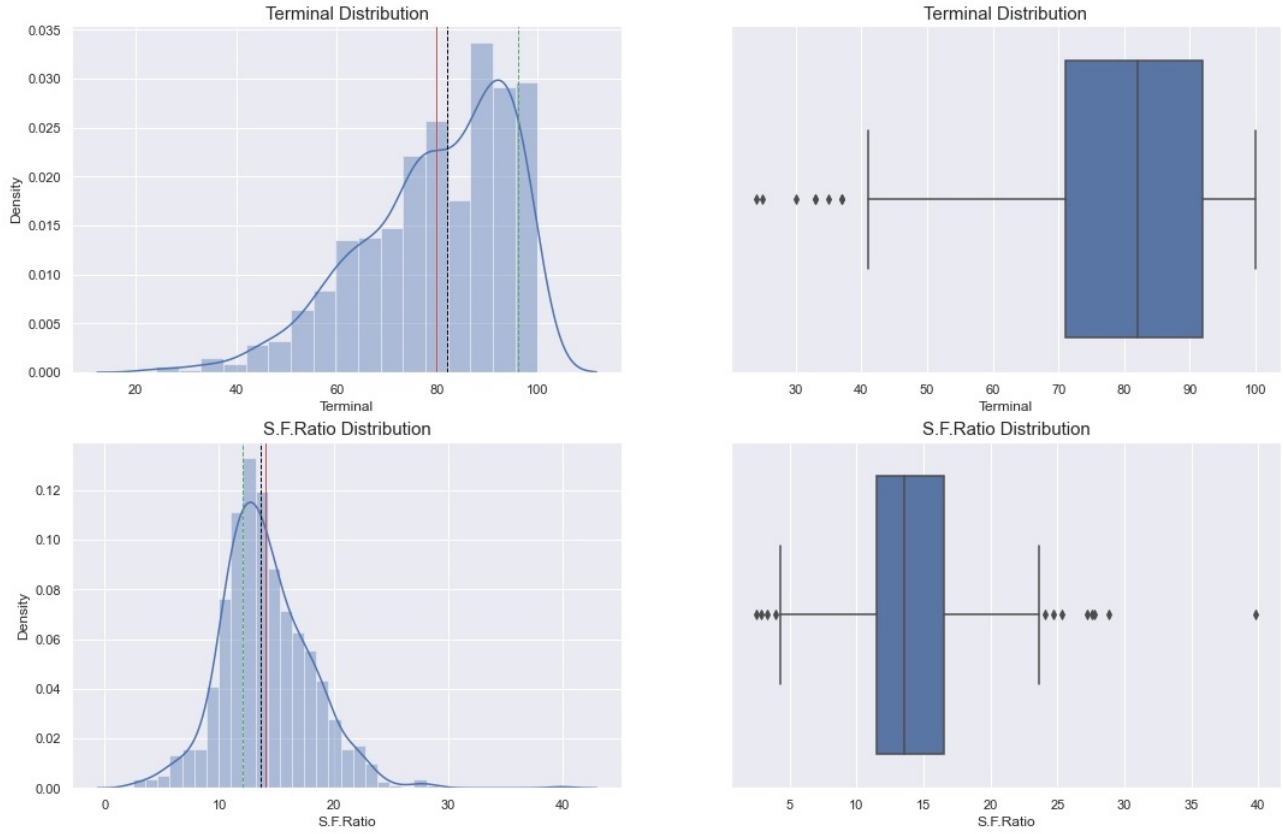


Fig 2.h.2 : Dist and Box Plot - Percent of Faculty with Terminal Degree and Student-Faculty ratio

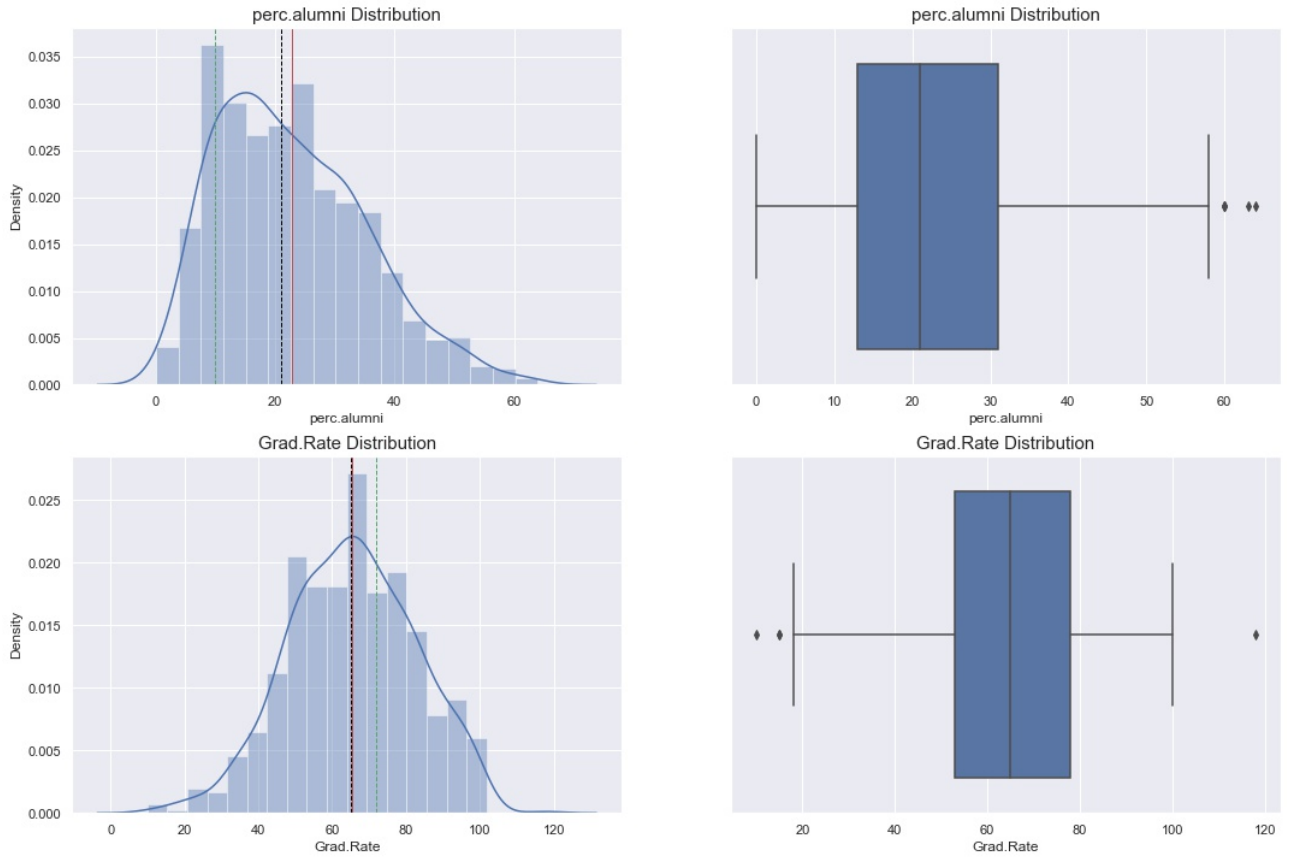


Fig 2.i : Dist and Box Plot - Percent of Alumni who Donate and Grad Rate

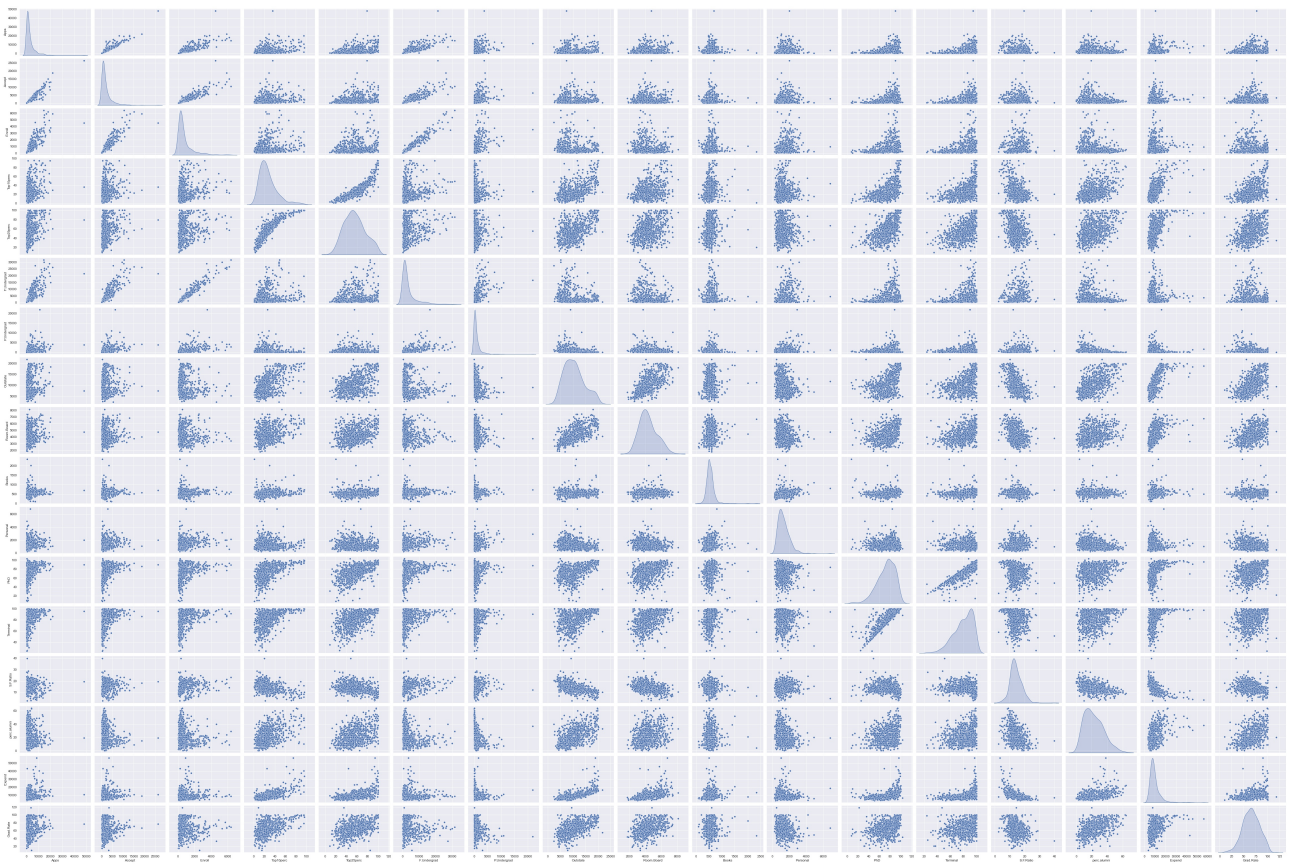

Fig 2.j : Pair Plot of All 17 Numeric Variables

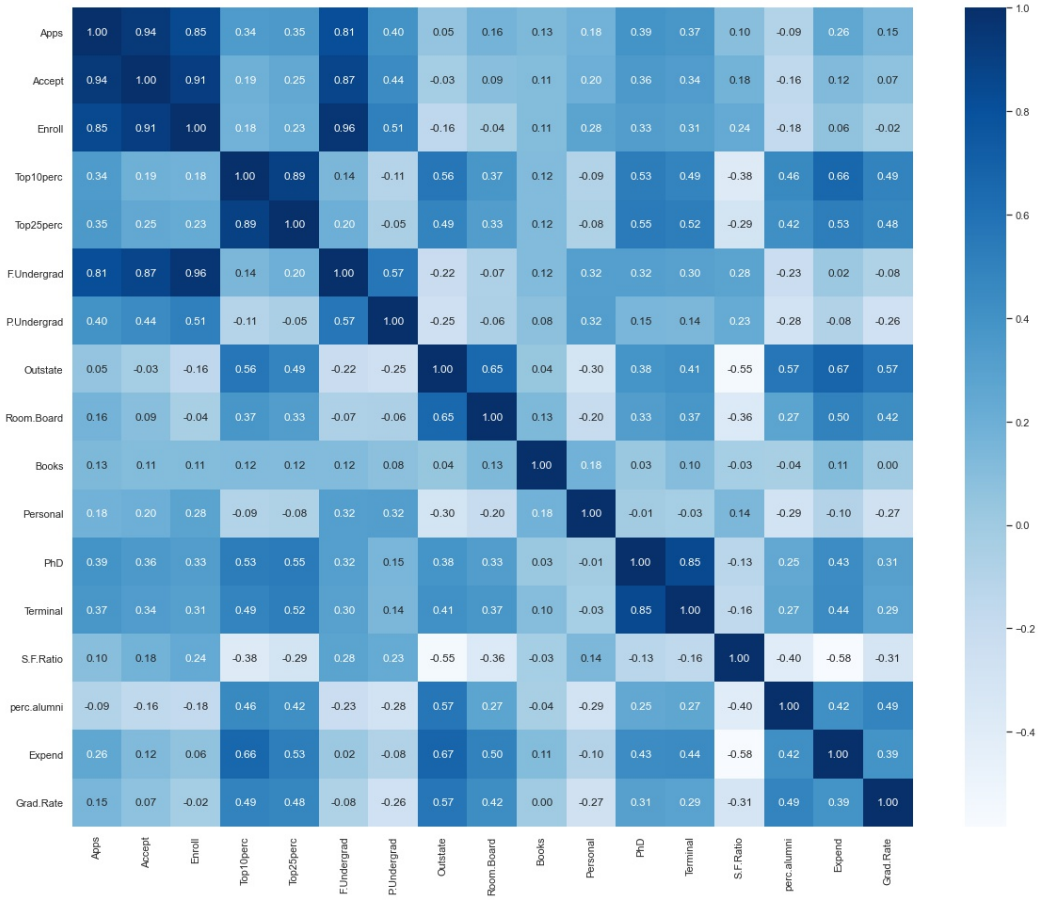


Fig 2.k : Correlation Heatmap

## UNIVARIATE ANALYSIS

1. In the DistPlots above, Mean, Median and Mode are denoted by Red, Black and Blue vertical lines

2. Lets group the 17 Numerical variables by association in 6 groups

   **GROUP 1 - APPLICATIONS**

   Apps --------> Mean = 3001.64, Median = 1558.0, CV = 128.94

   Accept --------> Mean = 2018.8, Median = 1110.0, CV = 121.41

   Enroll --------> Mean = 779.97, Median = 434.0, CV = 119.13

3. **GROUP 2 - QUALITY OF ENROLMENTS**

   Top10perc --------> Mean = 27.56, Median = 23.0, CV = 64.01

   Top25perc --------> Mean = 55.8, Median = 54.0, CV = 35.49

4. **GROUP 3 - ENROLLED STUDENTS' TYPE**

F.Undergrad --------> Mean = 3699.91, Median = 1707.0, CV = 131.1

P.Undergrad --------> Mean = 855.3, Median = 353.0, **CV = 178.0**

Outstate --------> Mean = 10440.67, Median = 9990.0, CV = 38.53

5. **GROUP 4 - EXPENSES**

Room.Board --------> Mean = 4357.53, Median = 4200.0, CV = 25.17

Books --------> Mean = 549.38, Median = 500.0, CV = 30.05

Personal --------> Mean = 1340.64, Median = 1200.0, CV = 50.5

Expend --------> Mean = 9660.17, Median = 8377.0, CV = 54.05

6. **GROUP 5 - FACULTY AND STUDENT FACULTY RATIO**

PhD --------> Mean = 72.66, Median = 75.0, CV = 22.47

Terminal --------> Mean = 79.7, Median = 82.0, **CV = 18.47**

S.F.Ratio --------> Mean = 14.09, Median = 13.6, CV = 28.09

7. **GROUP 6 - OTHERS - ALUMNI WHO DONATE AND GRAD RATE**

perc.alumni --------> Mean = 22.74, Median = 21.0, CV = 54.48

Grad.Rate --------> Mean = 65.46, Median = 65.0, CV = 26.24

8. We see that Percentage Coefficient of Variation (CV), the max CV = 178 for P.Undergrad

——-> implies that No. Of Part-time Undergrads vary largely from college to college and there is no consistency there

9. CV for Terminal = 18.47 and CV for PhD = 22.47, they are the 2 minimum CV of all the variables

——-> implies that percentage of Faculties with Terminal and PhD degrees is quite similar in all colleges with respect to their size and needs and there is consistency here

10. We notice that **values of skewness are negative** for Ph.D, Terminal and Grad.Rate

——-> implies that there are a few colleges with very less number of faculties with Ph.D and Terminal degrees and some colleges with very less rates of graduation

11. Consolidating the above points 8, 9 and 10 - though there are a few colleges having very less percentage of faculties with Ph.D and Terminal degrees, those colleges might be smaller with less intake too.

**MULTIVARIATE ANALYSIS**

1. There is a high correlation among No. Of Applications, Acceptance and Enrolments
2. Also, high correlation between No. Of Applications, Acceptances, Enrolments with Full Time Graduate Enrolments
3. We see high correlation between PhD and Terminal Degree.
4. We note that Outstate candidates spend more on Room Boarding - good correlation between these two
5. Also by its correlation, we see that Top10perc spend more on Instructional Expense. That is, we can safely say that higher rated colleges preferred by Top 10 Percentiles have more Tuition fees.

## [Q 2.2] Scale the variables and write the inference for using the type of scaling function for this case study.

| Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate |
|---|---|---|---|---|---|---|---|
| -0.376 | -0.338 | 0.106 | -0.247 | -0.192 | -0.019 | -0.166 | -0.746 |
| -0.159 | 0.117 | -0.260 | -0.696 | -1.354 | -0.094 | 0.798 | 0.458 |
| -0.472 | -0.427 | -0.569 | -0.311 | -0.293 | -0.704 | -0.778 | 0.201 |
| -0.890 | -0.918 | -0.919 | 2.129 | 1.678 | -0.899 | -0.828 | 0.627 |
| -0.983 | -1.051 | -1.063 | -0.696 | -0.596 | -0.996 | 0.298 | -0.717 |

| Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|---|---|---|---|---|
| -0.968 | -0.777 | 1.438 | -0.174 | -0.123 | 1.071 | -0.870 | -0.631 | -0.319 |
| 1.922 | 1.829 | 0.289 | -2.746 | -2.785 | -0.490 | -0.546 | 0.396 | -0.553 |
| -0.555 | -1.211 | -0.261 | -1.240 | -0.953 | -0.304 | 0.591 | -0.132 | -0.669 |
| 1.004 | -0.777 | -0.737 | 1.206 | 1.190 | -1.679 | 1.159 | 2.288 | -0.378 |
| -0.216 | 2.219 | 0.289 | 0.202 | -0.538 | -0.569 | -1.682 | 0.512 | -2.917 |

Table 2.4 : Education Data - ZSCORE (Standard Scaling) Applied

- Further, as we need to do PCA, we treat the Outliers and then do Scaling.
- Standardisation is very important for PCA as it is sensitive to Variance of the variables. Hence, it is very important that all variables have *comparative* scales.
- By Comparative Scales, we mean that all variables should be weighed in the same unit - which is almost never the case for any real data.
- If there are large differences between the ranges of some variables, those variables with larger ranges will dominate over those with small ranges
- For example, range (max - min) of variable 'Apps' = 48018, whereas for 'S.F.Ratio' is 37.3, hence 'Apps' variable will dominate over 'SF.Ratio' variable, which will lead to biased result.

  So, transforming the data to comparable scales can prevent this problem.
- Easiest and the most accepted form of Scaling is STANDARD SCALING a.k.a using ZSCORE, where we subtract the mean and divide by the Standard Deviation of each value of each variable

$$z_{value} = \frac{value - mean}{standard\ deviation}$$

- This pivots each variable about its mean and it is set to zero.
- After scaling, now, each value can be measured as the standard deviation distance from the mean on either side.
- Also, This is most widely used because with a spread of just 3 standard deviations on either side of mean, we are able to capture 99.7% of the data.

## [Q 2.3] Comment on the comparison between covariance and the correlation matrix after scaling.

- "Covariance" indicates the direction of the linear relationship between variables
- "Correlation" measures both the strength and direction of the linear relationship between two variables
- Correlation is a function of the covariance.

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)\ .\ SD(Y)}$$

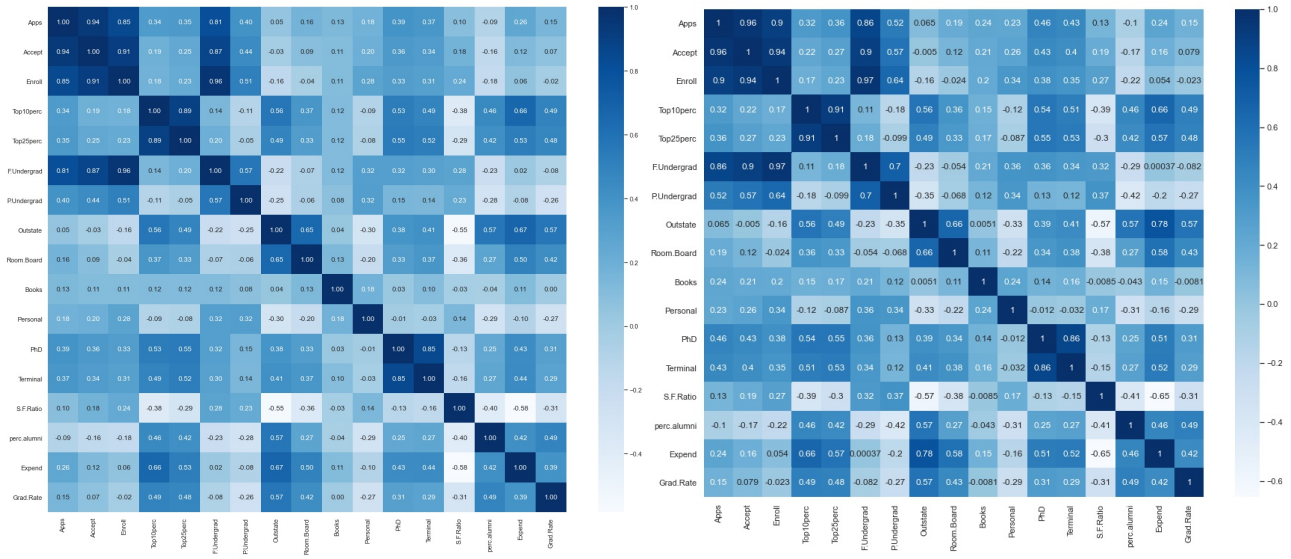

Fig 2.l : Correlation Heatmap - Before and After Scaling

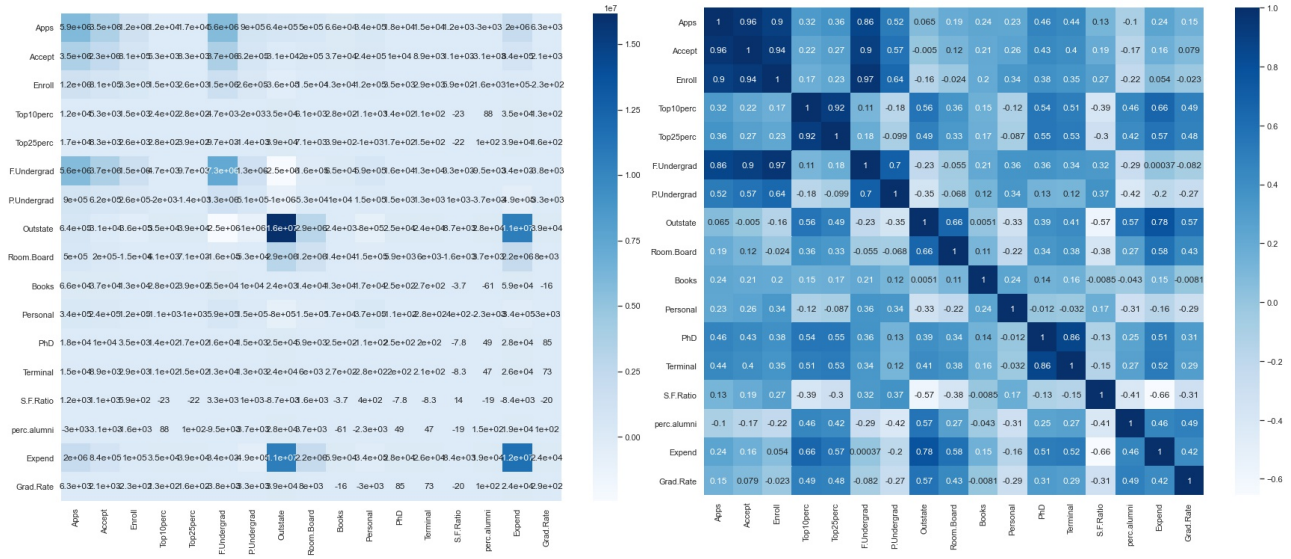

Fig 2.m : Covariance Heatmap - Before and After Scaling

- As seen in Fig 2.l - Correlation is not affected by Scaling. It remains the same before and after scaling.
- As seen in Fig 2.m - Covariance is largely affected by Scaling.
- Also evident from the above figures,

  CORRELATION MATRIX (Before And/Or After Scaling)

  = COVARIANCE MATRIX AFTER SCALING

## [Q 2.4] Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

- For this particular solution, Outlier treatment is not done
- Before Scaling, we can see that all variables have outliers except 'Top25perc'.
- Before Scaling, maximum variables have outliers on the right or the max side except 'Books', 'Ph.D', 'Terminal', 'S.F.Ratio' and 'Grad.Rate', which have it on the lower side

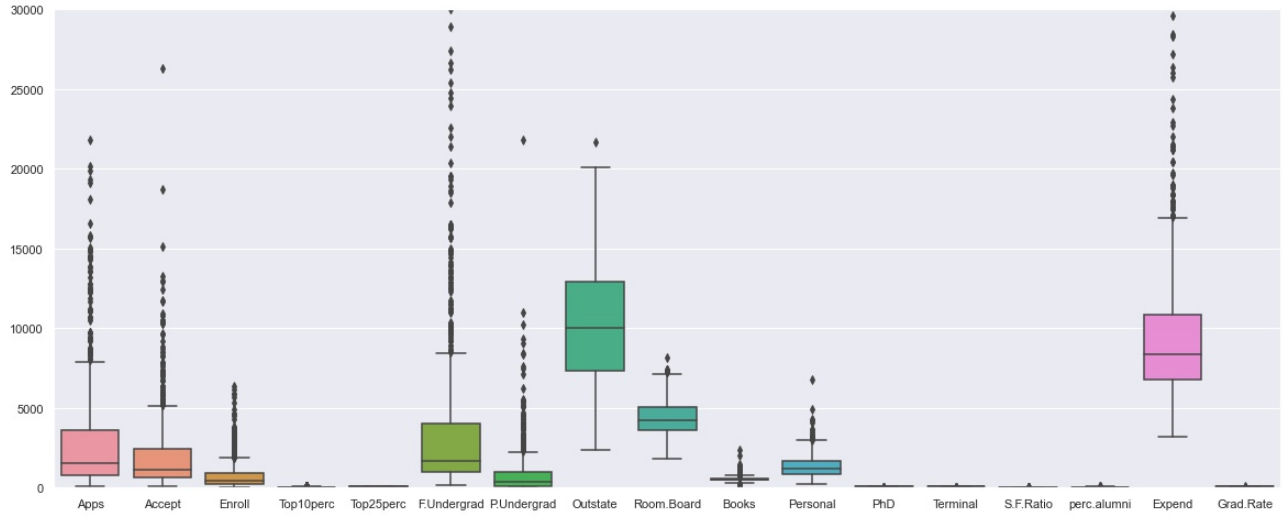


Fig 2.n : Boxplot of whole Education data - Before Scaling

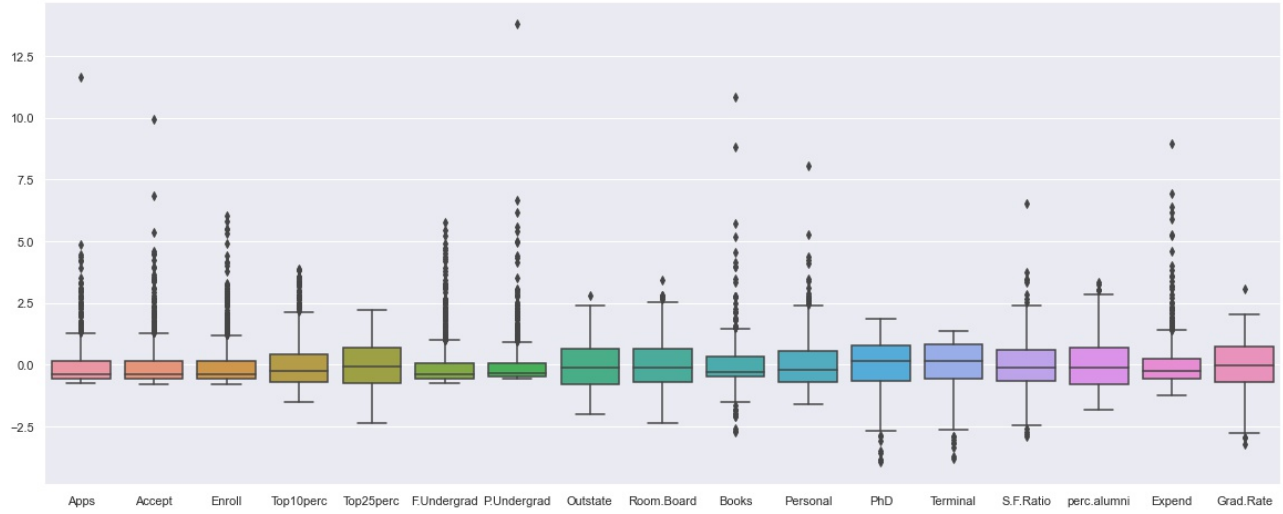


Fig 2.o : Boxplot of whole Education data - AFTER Z Scaling

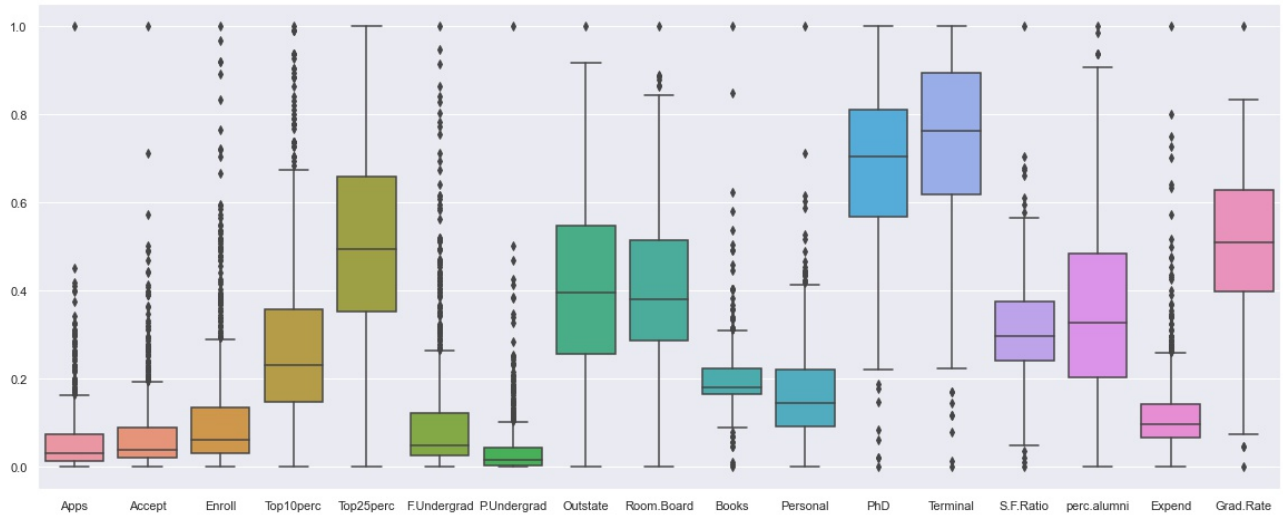


Fig 2.p : Boxplot of whole Education data - AFTER MinMax Scaling

- Before Scaling, we can see that all variables have outliers except 'Top25perc'.
- Before Scaling, maximum variables have outliers on the right or the max side except 'Books', 'Ph.D', 'Terminal', 'S.F.Ratio' and 'Grad.Rate', which have it on the lower side
- After Scaling - Both Standard and MinMax - we see that there has been NO effect on the existence of the Outliers.
- Outliers, along with all data, have simply been transformed to a different scale.
- It has been observed that in MinMax Scaling, data points are tightly packed, whereas in Standard scaling, it is unbounded
- Hence, we conclude that MinMax Scaling suppresses the effect of Outliers whereas Standard Scaling maintains useful information about Outliers.

## [Q 2.5] Build the covariance matrix and calculate the eigenvalues and the eigenvector.

- Eigen Values = [5.663, 4.895, 1.126, 1.004, 0.872, 0.766, 0.585, 0.545, 0.424, 0.381, 0.247, 0.022, 0.038, 0.147, 0.134, 0.099, 0.075]

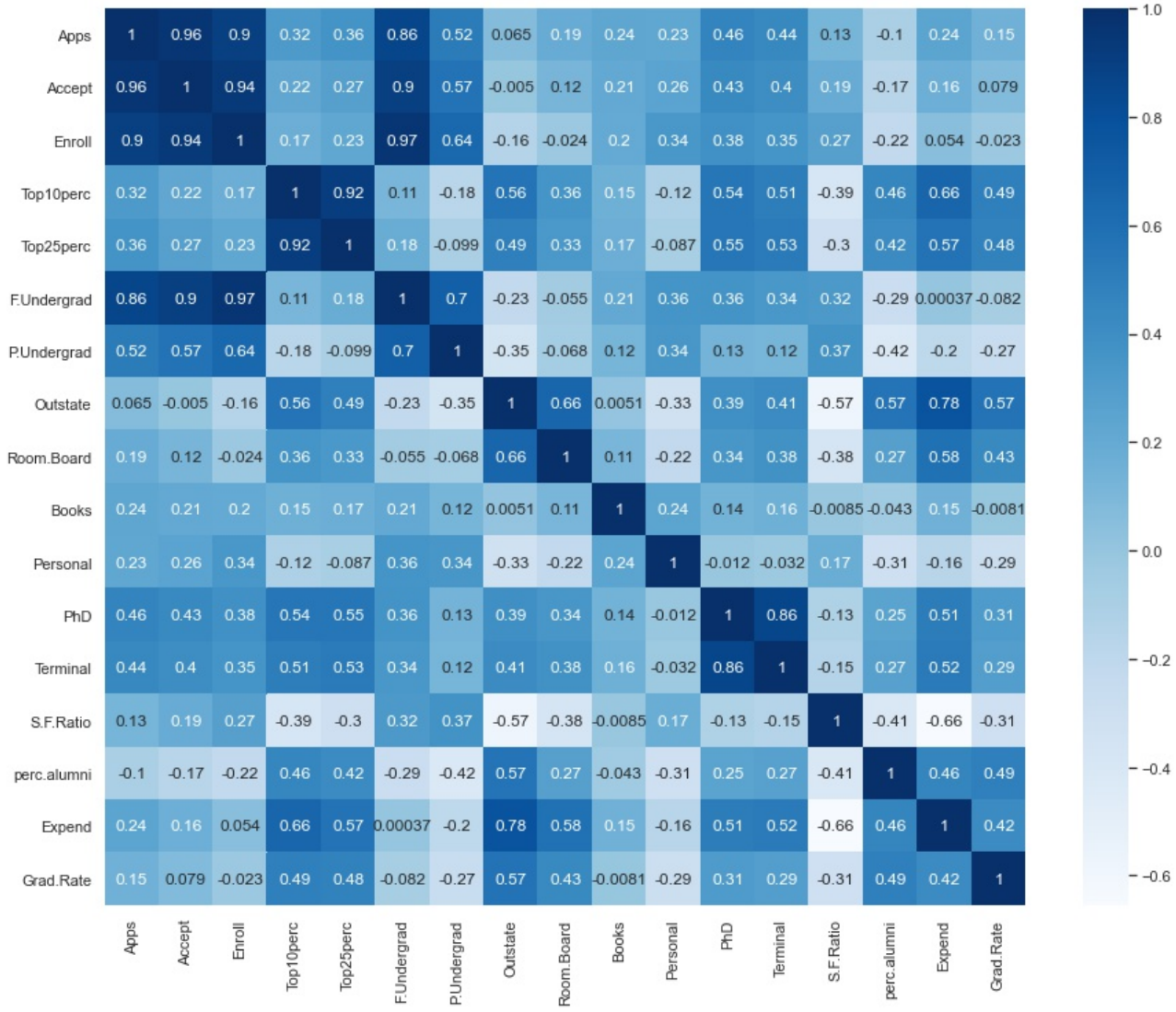


Fig 2.q : Covariance Matrix Heatmap - After Standard Scaling

- Eigen Vectors :
  - For Eigen value = 5.663

    Eigen Vector = [-0.262 0.314 0.081 -0.099 -0.22 0.002 -0.028 -0.09 0.131 -0.156 -0.086 0.182 -0.599 0.09 0.089 0.549 0.005]

  - For Eigen value = 4.895

    Eigen Vector = [-0.231 0.345 0.108 -0.118 -0.19 -0.017 -0.013 -0.138 0.142 -0.149 -0.043 -0.391 0.661 0.159 0.044 0.292 0.014]

  - For Eigen value = 1.126

    Eigen Vector = [-0.189 0.383 0.086 -0.009 -0.162 -0.068 -0.015 -0.144 0.051 -0.065. -0.044 0.717 0.233 -0.035 -0.062 -0.417 -0.05 ]

- For Eigen value = 1.004
  Eigen Vector = [-0.339 -0.099 -0.079  0.369 -0.157 -0.089 -0.257  0.29  -0.122
  -0.036. 0.002 -0.056  0.022 -0.039  0.07   0.009 -0.724]

- For Eigen value = 0.872
  Eigen Vector = [-0.335 -0.06  -0.051  0.417 -0.144 -0.028 -0.239  0.346 -0.194
  0.006. -0.102  0.02   0.032  0.146 -0.097 -0.011  0.655]

- For Eigen value = 0.766
  Eigen Vector = [-0.163  0.399  0.074 -0.014 -0.103 -0.052 -0.031 -0.109  0.001 -0.0
  -0.035 -0.543 -0.368 -0.134 -0.087 -0.571  0.025]

- For Eigen value = 0.585
  Eigen Vector = [-0.022  0.358  0.04  -0.225  0.096 -0.025 -0.01   0.124 -0.635
  0.546. 0.252  0.03   0.026  0.05   0.045  0.146 -0.04 ]

- For Eigen value = 0.545
  Eigen Vector = [-0.284 -0.252  0.015 -0.263 -0.037 -0.02   0.095  0.011 -0.008
  -0.232  0.593  0.001 -0.081  0.56   0.067 -0.212 -0.002]

- For Eigen value = 0.424
  Eigen Vector = [-0.244 -0.132 -0.021 -0.581  0.069  0.237  0.095  0.39  -0.221
  -0.255. -0.475  0.01   0.027 -0.107  0.018 -0.101 -0.028]

- For Eigen value = 0.381,
  Eigen Vector = [-0.097  0.094 -0.697  0.036 -0.035  0.639 -0.111 -0.24   0.021
  0.091. 0.044  0.004  0.01   0.052  0.035 -0.029 -0.008]

- For Eigen value = 0.247
  Eigen Vector = [ 0.035  0.232 -0.531  0.115  0.    -0.381  0.639  0.277  0.017 -0.128
  0.015 -0.011  0.005  0.009 -0.012  0.034  0.001]

- For Eigen value = 0.022
  Eigen Vector = [-0.326  0.055  0.081  0.147  0.551  0.003  0.089 -0.034  0.167
  0.101. -0.039  0.013  0.013 -0.072  0.703 -0.064  0.083]

- For Eigen value = 0.038
  Eigen Vector = [-0.323  0.043  0.059  0.089  0.59   0.035  0.092 -0.09   0.113
  0.086  -0.085  0.007 -0.018  0.164 -0.662  0.099 -0.113]

- For Eigen value = 0.147

  Eigen Vector = [ 0.163 0.26 0.274 0.259 0.143 0.469 0.153 0.243 -0.154 -0.471
  0.363 0.009 0.018 -0.24 -0.048 0.062 0.004]

- For Eigen value = 0.134

  Eigen Vector = [-0.187 -0.257 0.104 0.224 -0.128 0.013 0.391 -0.566 -0.539
  -0.148. -0.174 -0.024 -0. -0.049 0.036 0.028 -0.007]

- For Eigen value = 0.099

  Eigen Vector = [-0.329 -0.16 -0.184 -0.214 0.022 -0.232 -0.151 -0.119 0.024 -0.08
  0.394 0.011 0.056 -0.69 -0.127 0.129 0.145]

- For Eigen value = 0.075

  Eigen Vector = [-0.239 -0.168 0.245 0.036 -0.357 0.314 0.469 0.18 0.316
  0.488. 0.087 -0.003 0.015 -0.159 -0.063 -0.007 -0.003]

## [Q 2.6] Write the explicit form of the first PC (in terms of Eigen Vectors).

- Explicit form of first PC :

  (0.262 x Apps) + (0.231 x Accept) + (0.189 x Enroll) + (0.339 x Top10perc) + (0.335 x Top25perc) + (0.163 x F.Undergrad) + (0.022 x P.Undergrad) + (0.284 x Outstate) + (0.244 x Room.Board) + (0.097 x Books) + (-0.035 x Personal) + (0.326 x PhD) + (0.323 x Terminal) + (-0.163 x S.F.Ratio) + (0.187 x perc.alumni) + (0.329 x Expend) + (0.239 x Grad.Rate)

## [Q 2.7] Discuss 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? Perform PCA and export the data of the Principal Component scores into a data frame.

- **Cumulative values of Eigen Values as percentage of total are -**

  [ 33.266, 62.021, 68.639, 74.537, 79.661, 84.159, 87.596, 90.794, 93.282, 95.521, 96.972, 97.837, 98.626, 99.207, 99.646, 99.868, 100. ]

- The above values indicate the cumulative percentage contribution by each Eigen value to the Total.
- For example, contribution of first Eigen value is 33.266% whereas contribution of first 5 Eigen values is 79.66% and first 8 is 90.79% and so on.
- In short, they indicate the amount of information contained by the Principal Components (PC) - like PC1 contains 33.26% information of all PCs and First 5 PCs contain 79.66% of total information and so on.

- **Optimum number of Principal Components -**
  - There are few rules to decide the optimum number of Principal Components -
    - Eigen value Rule
    - Proportion of Variance Explained Rule
    - SCREE Plot Elbow Rule.
  - **Eigen Value Rule** - This rule states that a Principal Component (PC) considered should have variance of at-least 1. Hence by this rule, **we take all PCs with Eigen Values greater than or equal to 1**
    - **In Education data example, by this rule, we'll take 4 PCs**
  - **Proportion of Variance Explained Rule** - The *proportion of variance explained* (Percentage Cumulative Eigen Values) identifies the optimal number of PCs to keep based on the total variability that we would like to account for
    - So, in our example, If we decide to retain at-least 80% of the variability, then we should choose first 5 PCs and if we decide to retain 90% then choose first 8 PCs
  - **SCREE Plot Elbow Rule** - This rule looks for the "elbow" in the SCREE plot. The Rule says that we select all components after the elbow and just before the line flattens out.
    - So, in our example, by this rule, we choose 4 PCs - refer the figure 2.r.

- **Eigen Vectors - What do they indicate?**
  - Eigen Values and Eigen Vectors are calculated with respect to the correlation matrix of the features (or covariance matrix of scaled features).

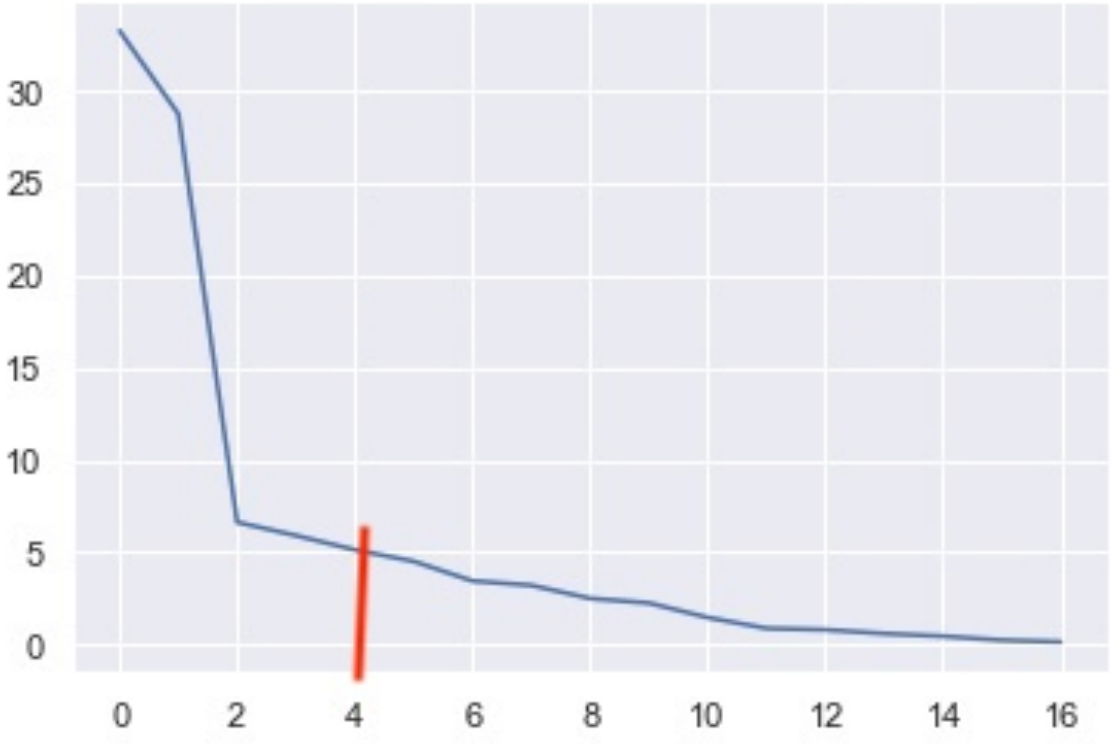


Fig 2.r : SCREE Plot - PCA of Education Data

- Eigen Values and Eigen Vectors are used to determine the Principal Components of the data.
- Eigen Values and Eigen Vectors always come in a pair, that is, every Eigen Vector has an Eigen Value. And their number is equal to the number of dimensions of the data (Number of Numeric variables/features of the data).
- **Eigen Vectors of the Covariance matrix are the directions of the axes where there is the most variance (most information) and these are called as the Principal Components.**
- Amount of Variance or Unit of Variance explained by these PCs is called Eigen Values

- **PCA and Principal Component Scores**
  - The following table shows all PCs and also simultaneously Eigen Vectors.
  - Columns show Principal Components and Rows show Eigen Vectors.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|
| EV1 | 0.262 | 0.314 | -0.081 | 0.099 | 0.22 | 0.002 | -0.028 | -0.09 | -0.131 |
| EV2 | 0.231 | 0.345 | -0.108 | 0.118 | 0.19 | -0.017 | -0.013 | -0.138 | -0.142 |
| EV3 | 0.189 | 0.383 | -0.086 | 0.009 | 0.162 | -0.068 | -0.015 | -0.144 | -0.051 |
| EV4 | 0.339 | -0.099 | 0.079 | -0.369 | 0.157 | -0.089 | -0.257 | 0.29 | 0.122 |
| EV5 | 0.335 | -0.06 | 0.051 | -0.417 | 0.144 | -0.028 | -0.239 | 0.346 | 0.194 |
| EV6 | 0.163 | 0.399 | -0.074 | 0.014 | 0.103 | -0.052 | -0.031 | -0.109 | -0.001 |
| EV7 | 0.022 | 0.358 | -0.04 | 0.225 | -0.096 | -0.025 | -0.01 | 0.124 | 0.635 |
| EV8 | 0.284 | -0.252 | -0.015 | 0.263 | 0.037 | -0.02 | 0.095 | 0.011 | 0.008 |
| EV9 | 0.244 | -0.132 | 0.021 | 0.581 | -0.069 | 0.237 | 0.095 | 0.39 | 0.221 |
| EV10 | 0.097 | 0.094 | 0.697 | -0.036 | 0.035 | 0.639 | -0.111 | -0.24 | -0.021 |
| EV11 | -0.035 | 0.232 | 0.531 | -0.115 | 0.0 | -0.381 | 0.639 | 0.277 | -0.017 |
| EV12 | 0.326 | 0.055 | -0.081 | -0.147 | -0.551 | 0.003 | 0.089 | -0.034 | -0.167 |
| EV13 | 0.323 | 0.043 | -0.059 | -0.089 | -0.59 | 0.035 | 0.092 | -0.09 | -0.113 |
| EV14 | -0.163 | 0.26 | -0.274 | -0.259 | -0.143 | 0.469 | 0.153 | 0.243 | 0.154 |
| EV15 | 0.187 | -0.257 | -0.104 | -0.224 | 0.128 | 0.013 | 0.391 | -0.566 | 0.539 |
| EV16 | 0.329 | -0.16 | 0.184 | 0.214 | -0.022 | -0.232 | -0.151 | -0.119 | -0.024 |
| EV17 | 0.239 | -0.168 | -0.245 | -0.036 | 0.357 | 0.314 | 0.469 | 0.18 | -0.316 |

| PC10 | PC11 | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 |
|---|---|---|---|---|---|---|---|
| -0.156 | -0.086 | -0.09 | -0.089 | -0.549 | 0.005 | 0.599 | -0.182 |
| -0.149 | -0.043 | -0.159 | -0.044 | -0.292 | 0.014 | -0.661 | 0.391 |
| -0.065 | -0.044 | 0.035 | 0.062 | 0.417 | -0.05 | -0.233 | -0.717 |
| -0.036 | 0.002 | 0.039 | -0.07 | -0.009 | -0.724 | -0.022 | 0.056 |
| 0.006 | -0.102 | -0.146 | 0.097 | 0.011 | 0.655 | -0.032 | -0.02 |
| 0.0 | -0.035 | 0.134 | 0.087 | 0.571 | 0.025 | 0.368 | 0.543 |
| 0.546 | 0.252 | -0.05 | -0.045 | -0.146 | -0.04 | -0.026 | -0.03 |
| -0.232 | 0.593 | -0.56 | -0.067 | 0.212 | -0.002 | 0.081 | -0.001 |
| -0.255 | -0.475 | 0.107 | -0.018 | 0.101 | -0.028 | -0.027 | -0.01 |
| 0.091 | 0.044 | -0.052 | -0.035 | 0.029 | -0.008 | -0.01 | -0.004 |
| -0.128 | 0.015 | -0.009 | 0.012 | -0.034 | 0.001 | -0.005 | 0.011 |
| 0.101 | -0.039 | 0.072 | -0.703 | 0.064 | 0.083 | -0.013 | -0.013 |
| 0.086 | -0.085 | -0.164 | 0.662 | -0.099 | -0.113 | 0.018 | -0.007 |
| -0.471 | 0.363 | 0.24 | 0.048 | -0.062 | 0.004 | -0.018 | -0.009 |
| -0.148 | -0.174 | 0.049 | -0.036 | -0.028 | -0.007 | 0.0 | 0.024 |
| -0.08 | 0.394 | 0.69 | 0.127 | -0.129 | 0.145 | -0.056 | -0.011 |
| 0.488 | 0.087 | 0.159 | 0.063 | 0.007 | -0.003 | -0.015 | 0.003 |

## [Q 2.8] Mention the business implication of using the Principal Component Analysis for this case study.

- Principal Component Analysis (PCA) is used to seek the most accurate data representation in a lower dimensional space
- Reducing the dimension of the feature space is called 'Dimensionality Reduction'
- We are able to analyse the data more effectively - when we drop the "least important" variables while still retaining the maximum valuable information in the dataset
- Our Education data, has 17 Numerical Variables which talks about the Admission Cycle of 777 different colleges from No. Of Applications, Acceptances, Enrolments to various Expenses incurred by Students to Quality of faculty to Graduation Rate.
- Here, 80% of total information is captured by 5 Components.
- So, we reduce dimension of feature space from 17 to 5

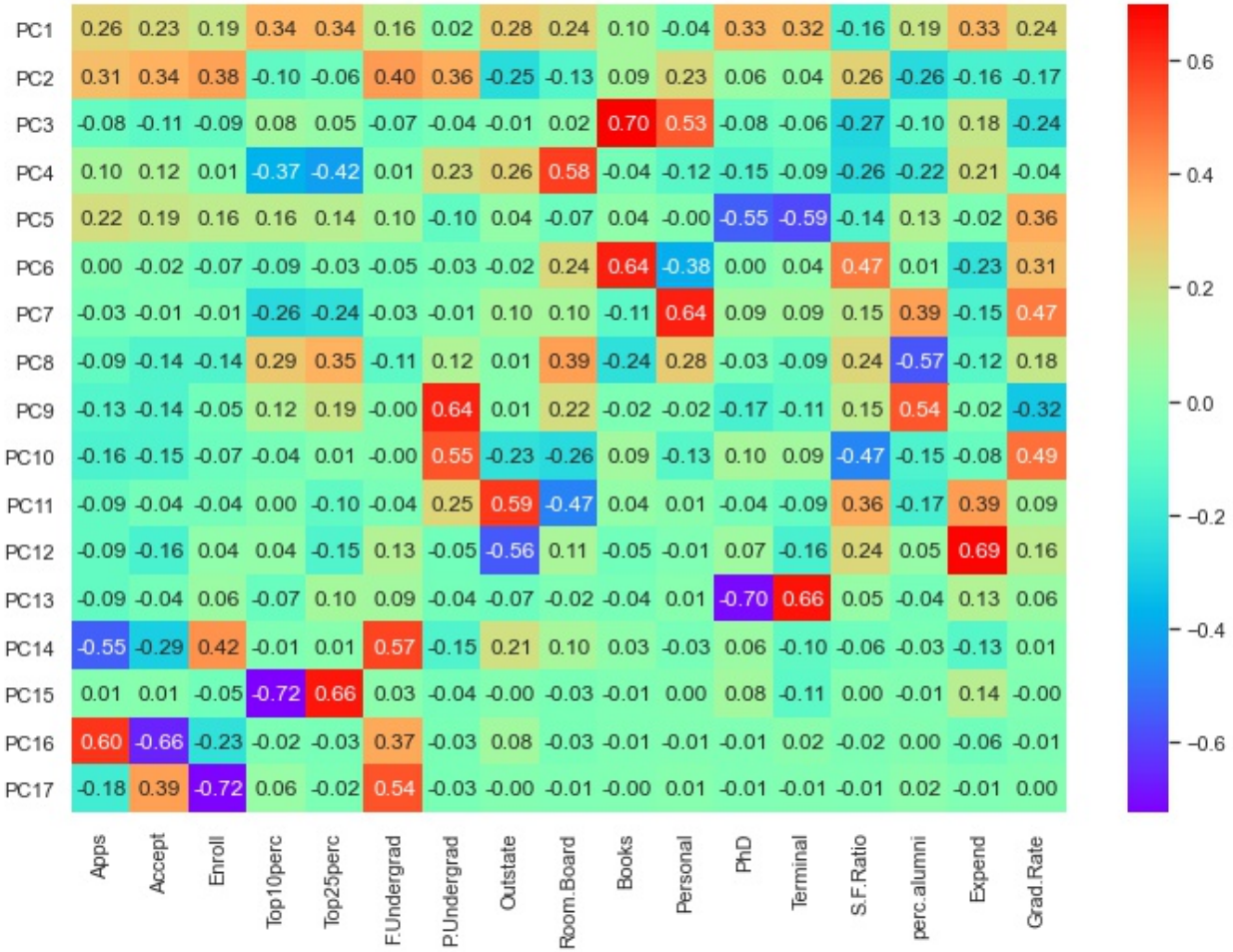

Fig 2.s : Component vs Features Heatmap

- These 5 components, not only retain maximum information but are also independent of each other, thus making it highly effective for most modelling.
- PC1 - looks more related to Top10perc, Top25perc, PhD, Terminal and Expend
  - we can label PC1 as 'Academic Merit and Tuition'
- PC2 - looks more related to Apps, Accept, Enrol, F.Undergrad and P.Undergrad
  - we can label it as 'Application Matrix'
- PC3 - has high correlation with Books
  - we can label it as 'Expenditure on Books'
- PC4 - has good correlation with Room Board
  - we can label it as 'Lodging Expenses'
- PC5 - has good correlation with PhD and Terminal
  - we can label it as 'Faculty Calibre'

———————————————-