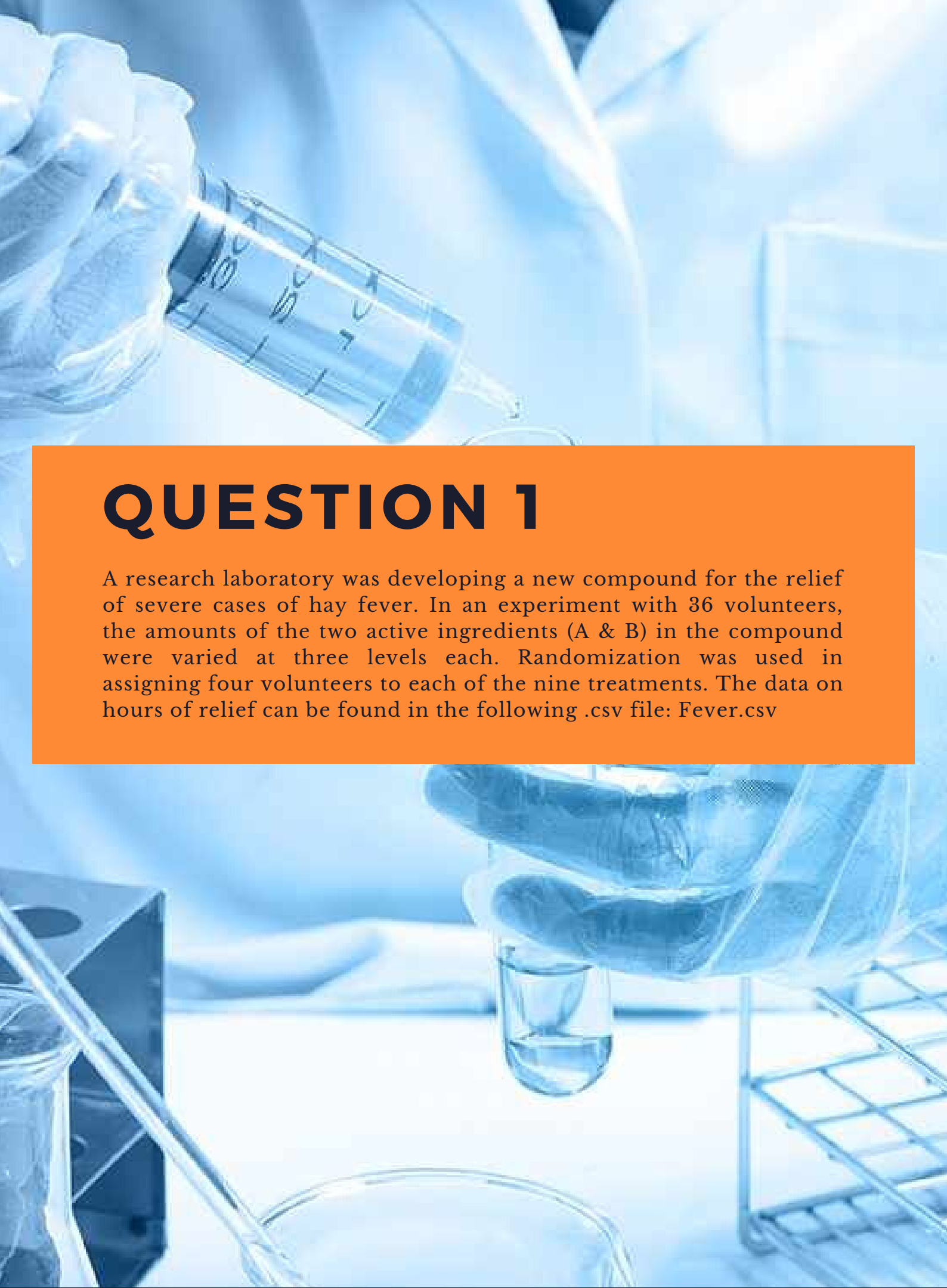
# ADVANCED STATISTICS PROJECT


**PREPARED BY:**

Jotinder Singh Matta


**BATCH:-**

PGP DSBA March20A



# QUESTION 1

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. Randomization 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

# 1.1) STATE THE NULL AND ALTERNATE HYPOTHESIS FOR CONDUCTING ONE-WAY ANOVA FOR BOTH THE VARIABLES 'A' AND 'B' INDIVIDUALLY.

## VARIABLE A

**Null Hypothesis (H0)** --> All the means are equal --> i.e. Different levels of Compound A (1,2,3) have no effect on the relief variable.

**H0 --> $\mu 1=\mu 2=\mu 3$**

**Alternate Hypothesis (Ha)** --> At least one of the means is not equal --> i.e. Different levels of Compound A (1,2,3) have effect on the relief variable

**Ha --> $\mu 1\neq\mu 2\neq\mu 3$**

## VARIABLE B

**Null Hypothesis (H0)** --> All the means are equal --> i.e. Different levels of Compound B (1,2,3) have no effect on the relief variable.

**H0 --> $\mu 1=\mu 2=\mu 3$**

**Alternate Hypothesis (Ha)** --> At least one of the means is not equal --> i.e. Different levels of Compound B (1,2,3) have effect on the relief variable

**Ha --> $\mu 1\neq\mu 2\neq\mu 3$**

# 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.

## HYPOTHESIS

**Null Hypothesis (H0)** --> All the means are equal --> i.e. Different levels of Compound A (1,2,3) have no effect on the relief variable.

**H0 --> $\mu 1 = \mu 2 = \mu 3$**

**Alternate Hypothesis (Ha)** --> At least one of the means is not equal --> i.e. Different levels of Compound A (1,2,3) have effect on the relief variable

**Ha --> $\mu 1 \neq \mu 2 \neq \mu 3$**

## ANOVA OUTPUT

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(A)** | 2.0 | 220.02 | 110.010000 | 23.465387 | 4.578242e-07 |
| **Residual** | 33.0 | 154.71 | 4.688182 | NaN | NaN |

## CONCLUSION

Since p-value is less than alpha (0.05) here, we **reject the null hypothesis** and conclude that difference between some of the means are statistically significant. Or in other words different levels of Compound A (i.e. 1,2,3) have a statistically significant impact on the relief variable.

# 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

## HYPOTHESIS

**Null Hypothesis (H0)** --> All the means are equal --> i.e. Different levels of Compound B (1,2,3) have no effect on the relief variable.

**H0 --> $\mu 1=\mu 2=\mu 3$**

**Alternate Hypothesis (Ha)** --> At least one of the means is not equal --> i.e. Different levels of Compound B (1,2,3) have effect on the relief variable

**Ha --> $\mu 1\neq\mu 2\neq\mu 3$**

## ANOVA OUTPUT

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(B) | 2.0 | 123.66 | 61.830000 | 8.126777 | 0.00135 |
| Residual | 33.0 | 251.07 | 7.608182 | NaN | NaN |

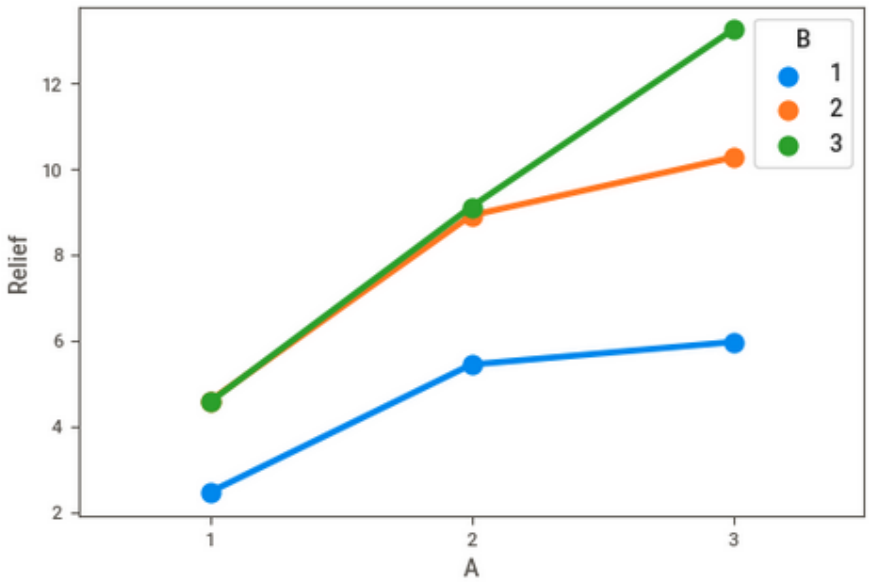
## CONCLUSION

Since p-value is less than alpha (0.05) here, we **reject the null hypothesis** and conclude that difference between some of the means are statistically significant. Or in other words different levels of Compound B (i.e. 1,2,3) have a statistically significant impact on the relief variable.

## 1.4) ANALYSE THE EFFECTS OF ONE VARIABLE ON ANOTHER WITH THE HELP OF AN INTERACTION PLOT.WHAT IS AN INTERACTION BETWEEN TWO TREATMENTS?

[HINT: USE THE 'POINTPLOT' FUNCTION FROM THE 'SEABORN' FUNCTION]

# INTERACTION PLOT




# ANOVA OUTPUT

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(A) | 2.0 | 220.020 | 110.010000 | 1827.858462 | 1.514043e-29 |
| C(B) | 2.0 | 123.660 | 61.830000 | 1027.329231 | 3.348751e-26 |
| C(A):C(B) | 4.0 | 29.425 | 7.356250 | 122.226923 | 6.972083e-17 |
| Residual | 27.0 | 1.625 | 0.060185 | NaN | NaN |

CONTINUED..

## CONCLUSION

- We can see there is some interaction between Variable A and Variable B
- We can see the two lines (Green and Orange) intersect each other, indicating Interaction between the two components.
- Based on ANOVA we can see there is statistically significant interaction between A and B.
- Therefore the effect of Compound A is dependent on the value of Compound B and vice-versa.

## WHAT IS AN INTERACTION BETWEEN TWO TREATMENTS?

- In statistics, an interaction may arise when considering the relationship among three or more variables, and describes a situation in which the effect of one causal variable on an outcome depends on the state of a second causal variable.
- If two variables of interest interact, the relationship between each of the interacting variables and a third "dependent variable" depends on the value of the other interacting variable.
- In our example since, since variables A & B interact, the relation between A and the relief variable depends on the value of variable B.
- Similarly, since variables A & B interact, the relation between B and the relief variable depends on the value of variable A.

# 1.5) PERFORM A TWO-WAY ANOVA BASED ON THE DIFFERENT INGREDIENTS (VARIABLE ‘A’ & ‘B’) WITH THE VARIABLE 'RELIEF' AND STATE YOUR RESULTS.

## HYPOTHESIS

- There are total 6 hypothesis here,
  - 3 Null and 3 Alternate hypothesis.
  - 2 each for variable A & B, and 2 for the interaction between A & B.

- **Hypothesis for Variable A.**
  - Null Hypothesis (H0) --> All the means are equal --> i.e. Different levels of Compound A (1,2,3) have no effect on the relief variable.
  - $\mu 1 = \mu 2 = \mu 3$
  - Alternate Hypothesis (Ha) --> At least one of the means is not equal --> i.e. Different levels of Compound A (1,2,3) do have an effect on the relief variable.
  - $\mu 1 \neq \mu 2 \neq \mu 3$

- **Hypothesis for Variable B.**
  - Null Hypothesis (H0) --> All the means are equal --> i.e. Different levels of Compound B (1,2,3) have no effect on the relief variable.
  - $\mu 1 = \mu 2 = \mu 3$
  - Alternate Hypothesis (Ha) --> At least one of the means is not equal --> i.e. Different levels of Compound B (1,2,3) do have an effect on the relief variable.
  - $\mu 1 \neq \mu 2 \neq \mu 3$

CONTINUED..

# HYPOTHESIS

- Hypothesis for Interaction between variable A & B.
  - Null Hypothesis (H0) --> There is no interaction between variable A & B.
  - Alternate Hypothesis (Ha) --> There is some interaction between variable A & B.

## 2-WAY ANOVA WITHOUT INTERACTION

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(A) | 2.0 | 220.02 | 110.010000 | 109.832850 | 8.514029e-15 |
| C(B) | 2.0 | 123.66 | 61.830000 | 61.730435 | 1.546749e-11 |
| Residual | 31.0 | 31.05 | 1.001613 | NaN | NaN |

## 2-WAY ANOVA WITH INTERACTION

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(A) | 2.0 | 220.020 | 110.010000 | 1827.858462 | 1.514043e-29 |
| C(B) | 2.0 | 123.660 | 61.830000 | 1027.329231 | 3.348751e-26 |
| C(A):C(B) | 4.0 | 29.425 | 7.356250 | 122.226923 | 6.972083e-17 |
| Residual | 27.0 | 1.625 | 0.060185 | NaN | NaN |

CONTINUED..

# ANOVA RESULTS

Based on the outcomes of the ANOVA tests performed above we can safely say that we are going to **reject all the null hypothesis and accept all three alternate hypothesis, since the p-value in each case is less than alpha (0.05) i.e.**

- Considering variable A, we can say it has a statistically significant impact on the relief variable.
- Considering variable B, we can say it has a statistically significant impact on the relief variable.
- Looking at ANOVA with interaction, we can say there is a statistically significant interaction between A & B.
  - Thus the relationship between A and relief variable depends on the value of the B variable.
  - and vice-versa.

## 1.6) MENTION THE BUSINESS IMPLICATIONS OF PERFORMING ANOVA FOR THIS PARTICULAR CASE STUDY.

# BUSINESS IMPLICATIONS

- After performing ANOVA for this case study and looking at the results, we can infer below:
    - Both Compounds A and B are independently critical to the relief given by the selected treatment.
    - Compound A has higher impact on relief variable compared to Compound B, looking at their respective p-values.
    - However Compounds A and B interact with each other, that means the amount of relief for a given treatment will depend on the levels selected for A and B respectively.
    - Choosing different levels of compound A, will bear implications on the effect of compound B on the overall relief.
    - Similarly different levels of compound B, will bear implications on the effect of compound A on the overall relief.

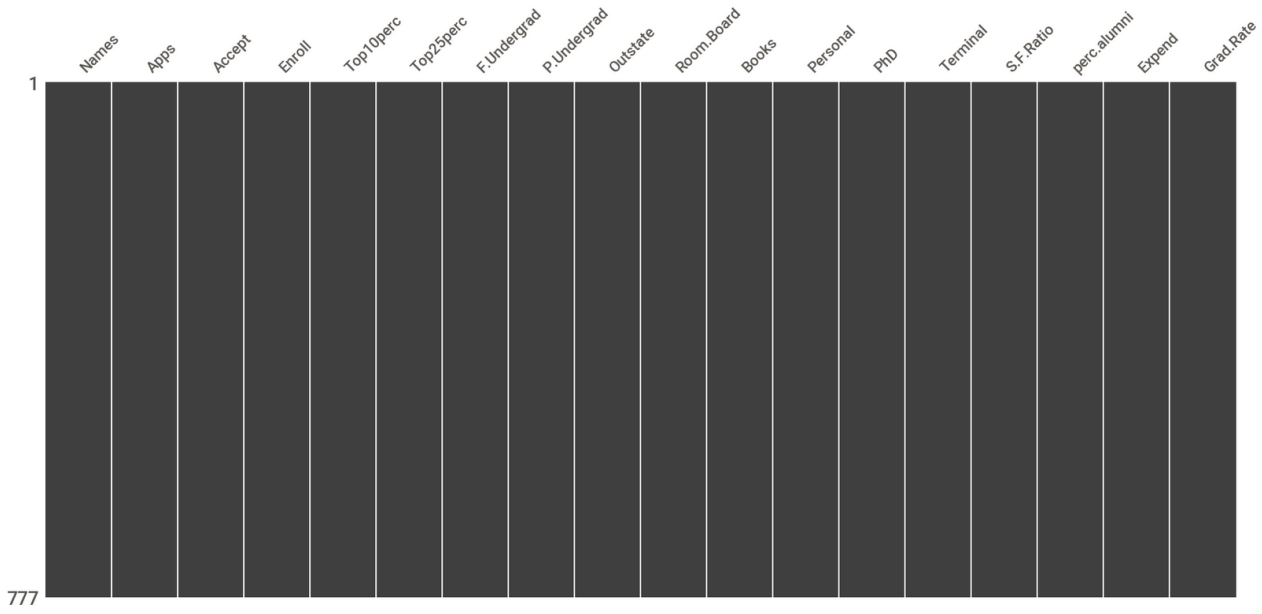
# QUESTION 2

The dataset Education - Post 12th Standard.csv is a dataset which 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.

2.1) PERFORM EXPLORATORY DATA ANALYSIS [BOTH UNIVARIATE AND MULTIVARIATE ANALYSIS TO BE PERFORMED]. THE INFERENCES DRAWN FROM THIS SHOULD BE PROPERLY DOCUMENTED.

# EXPLORATORY DATA ANALYSIS

## MISSING DATA ANALYSIS




- There are no missing values in the data set.
- We have used missingno library here to visualize the missing data to get a better sense of how the data looks.
- We also checked for missing data using .isna().sum() & .isnull().sum(), both of which returned 0.
- Hence we can see there are no missing values, hence missing data treatment is not required in this case.

CONTINUED..

# 5 POINT SUMMARY

| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate |
|---|---|---|---|---|---|---|---|---|
| count | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 |
| mean | 3001.638353 | 2018.804376 | 779.972973 | 27.558559 | 55.796654 | 3699.907336 | 855.298584 | 10440.669241 |
| std | 3870.201484 | 2451.113971 | 929.176190 | 17.640364 | 19.804778 | 4850.420531 | 1522.431887 | 4023.016484 |
| min | 81.000000 | 72.000000 | 35.000000 | 1.000000 | 9.000000 | 139.000000 | 1.000000 | 2340.000000 |
| 25% | 776.000000 | 604.000000 | 242.000000 | 15.000000 | 41.000000 | 992.000000 | 95.000000 | 7320.000000 |
| 50% | 1558.000000 | 1110.000000 | 434.000000 | 23.000000 | 54.000000 | 1707.000000 | 353.000000 | 9990.000000 |
| 75% | 3624.000000 | 2424.000000 | 902.000000 | 35.000000 | 69.000000 | 4005.000000 | 967.000000 | 12925.000000 |
| max | 48094.000000 | 26330.000000 | 6392.000000 | 96.000000 | 100.000000 | 31643.000000 | 21836.000000 | 21700.000000 |

| | Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|---|---|---|---|---|---|
| count | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.000000 | 777.00000 |
| mean | 4357.526384 | 549.380952 | 1340.642214 | 72.660232 | 79.702703 | 14.089704 | 22.743887 | 9660.171171 | 65.46332 |
| std | 1096.696416 | 165.105360 | 677.071454 | 16.328155 | 14.722359 | 3.958349 | 12.391801 | 5221.768440 | 17.17771 |
| min | 1780.000000 | 96.000000 | 250.000000 | 8.000000 | 24.000000 | 2.500000 | 0.000000 | 3186.000000 | 10.00000 |
| 25% | 3597.000000 | 470.000000 | 850.000000 | 62.000000 | 71.000000 | 11.500000 | 13.000000 | 6751.000000 | 53.00000 |
| 50% | 4200.000000 | 500.000000 | 1200.000000 | 75.000000 | 82.000000 | 13.600000 | 21.000000 | 8377.000000 | 65.00000 |
| 75% | 5050.000000 | 600.000000 | 1700.000000 | 85.000000 | 92.000000 | 16.500000 | 31.000000 | 10830.000000 | 78.00000 |
| max | 8124.000000 | 2340.000000 | 6800.000000 | 103.000000 | 100.000000 | 39.800000 | 64.000000 | 56233.000000 | 118.00000 |

- We performed 5 - point summary analysis to get better insights into the data.
- Looking at the summary we can see there is some skewness in the data.
- We further checked for skewness, to confirm our analysis.
- Interesting facts about the shape of the data were uncovered after checking for skewness.

# CONTINUED..

| | Amount of Skewness |
|---|---|
| Apps | 3.716557 |
| Accept | 3.411126 |
| Enroll | 2.685268 |
| Top10perc | 1.410487 |
| Top25perc | 0.258839 |
| F_Undergrad | 2.605416 |
| P_Undergrad | 5.681358 |
| Outstate | 0.508294 |

| | Amount of Skewness |
|---|---|
| Room_Board | 0.476434 |
| Books | 3.478293 |
| Personal | 1.739131 |
| PhD | -0.766686 |
| Terminal | -0.814965 |
| S_F_Ratio | 0.666146 |
| perc_alumni | 0.605719 |
| Expend | 3.452640 |
| Grad_Rate | -0.113558 |

- Looking at the skewness, we can make below conclusions.

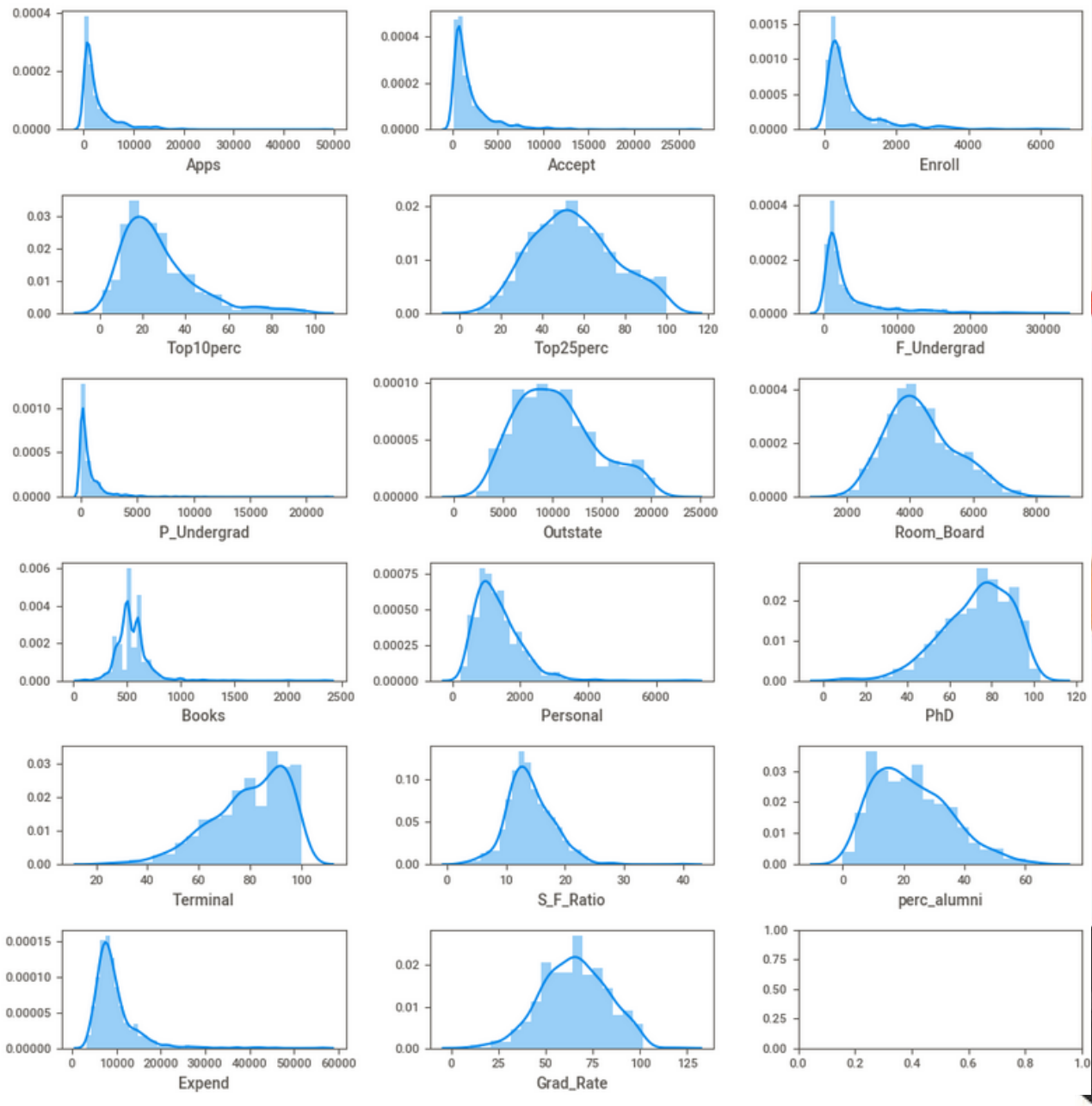
1. **Variables which are highly right skewed.**
   a. Apps
   b. Accept
   c. Enroll
   d. Top10perc
   e. F_Undergrad
   f. P_Undergrad
   g. Books
   h. Personal
   i. Expend

2. **Variables that are slightly right skewed.**
   a. Top25perc
   b. Outstate
   c. Room_Board
   d. S_F_Ratio
   e. perc_alumni

3. **Variables that are slightly left skewed**
   a. PhD
   b. Terminal
   c. Grad_Rate

CONTINUED..

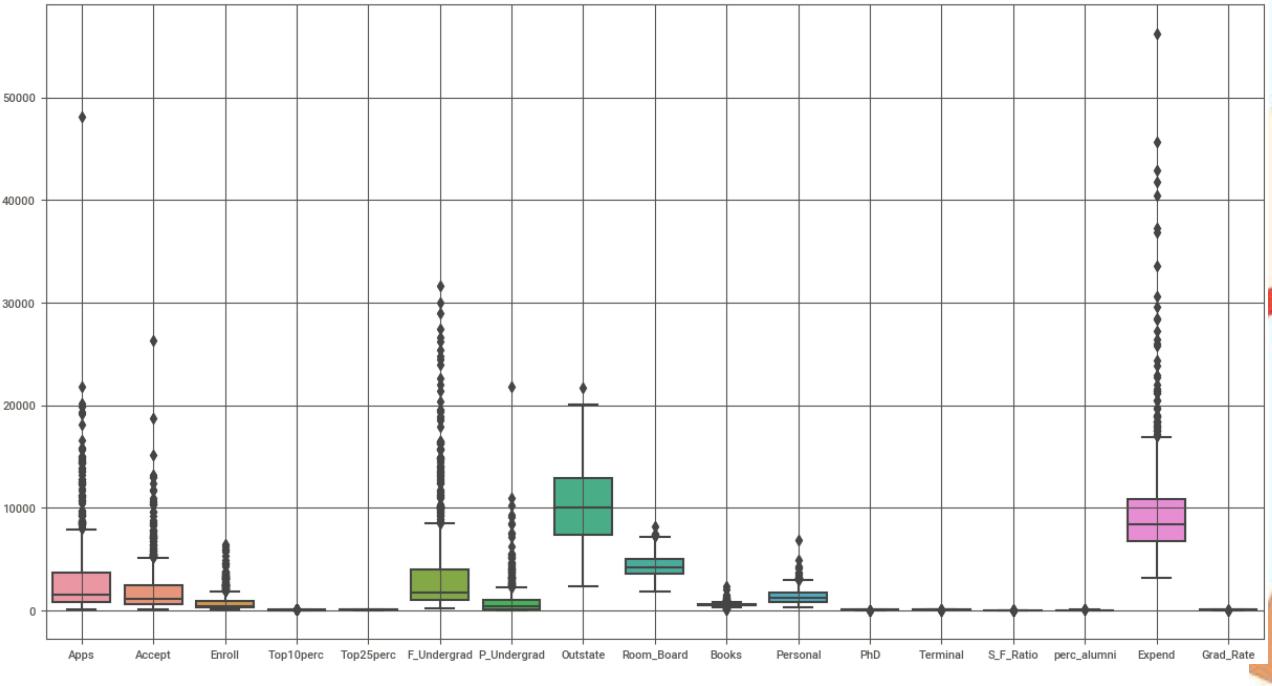
# UNIVARIATE ANALYSIS - DISTPLOT




Distplot further confirms our skewness analysis by visually plotting all the columns, so we can clearly see the left and right tails for the skewed columns.

CONTINUED..

# UNIVARIATE ANALYSIS - BOXPLOT

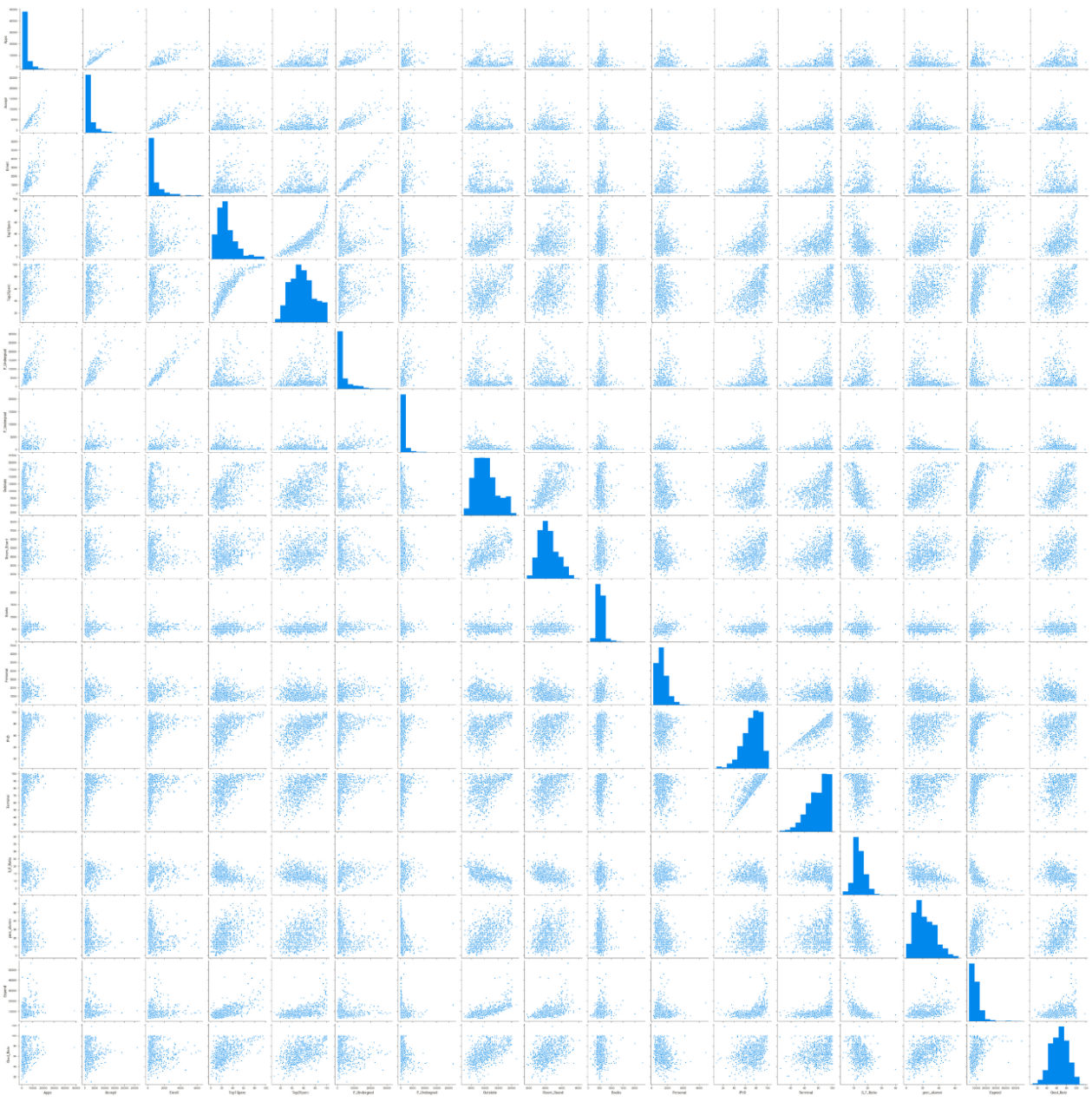



Box plot reveals a few insights for us.

- Skewness in data is further confirmed by this.
- We can see there are number of outliers present in the data set.
- Also we can see some of the columns are hardly visible while the others are clearly visible, indicating scaling problem with this data.
- In terms of absolute value, some of the columns have very high value compared to the others. E.g. Expend Vs. PhD

CONTINUED..

# BIVARIATE ANALYSIS - PAIRPLOT

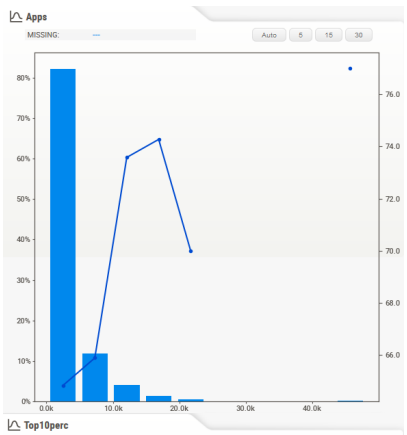


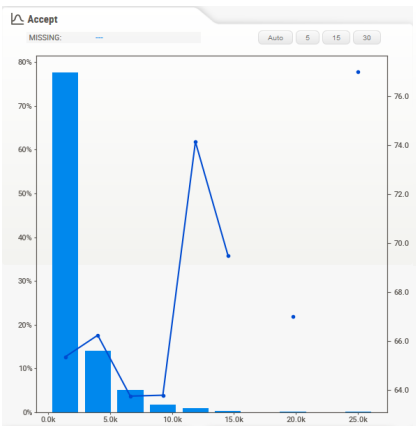
We can see various linear correlations present in the data by looking at this plot. E.g. F_Undergrad and Enroll, Terminal Vs. PhD etc.

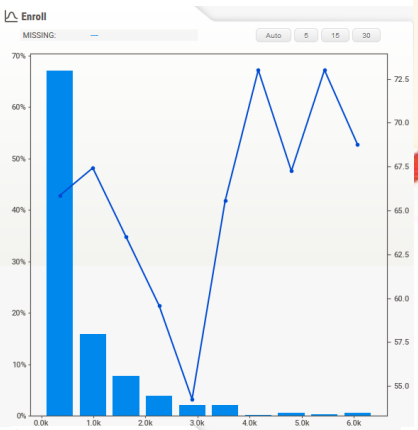
CONTINUED..

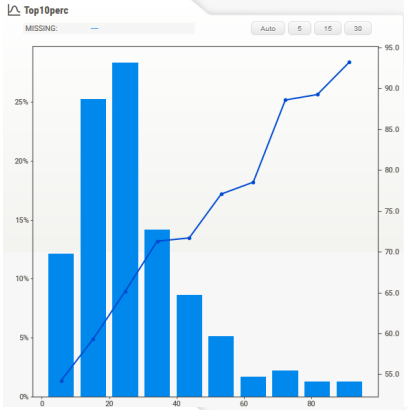
## BIVARIATE ANALYSIS - GRAD_RATE VS REST

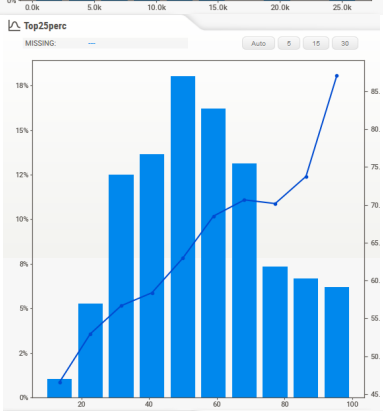
Here we have considered the Grad_Rate to be the dependent variable, as logically looking at the data set, we can see the based on various factors, finding out the Grad_Rate for a college would be the ultimate goal. We then compared all other variables w.r.t the Grad_Rate variable.

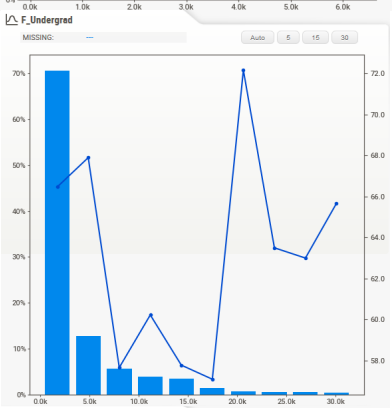


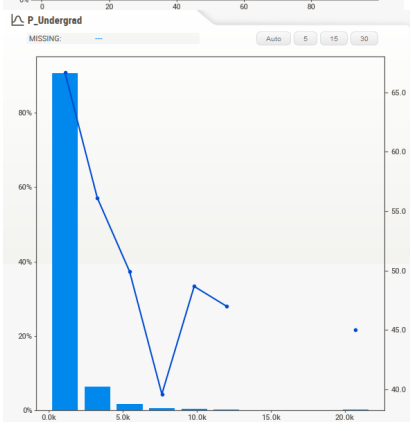



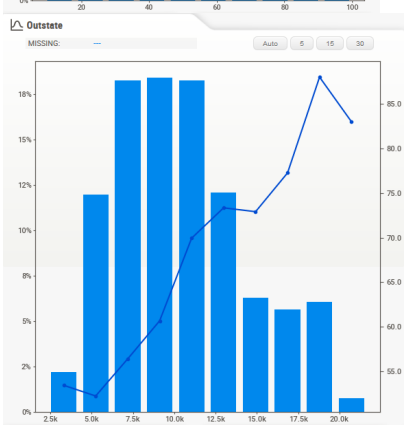



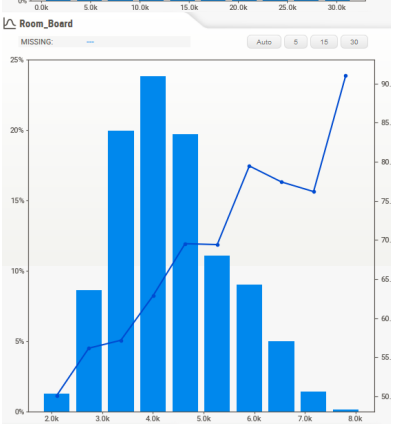



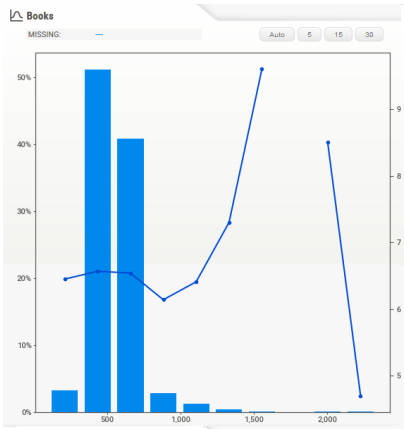



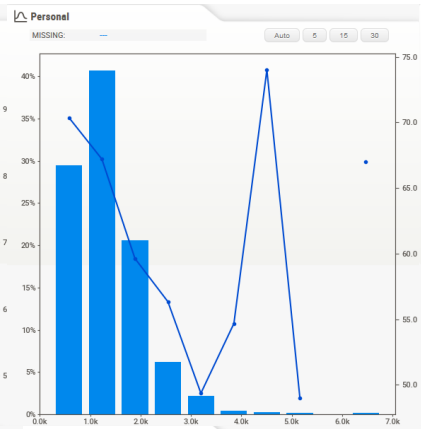



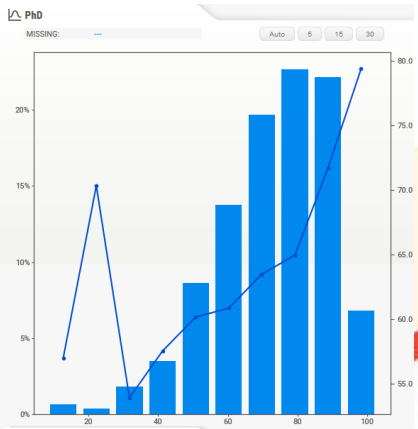



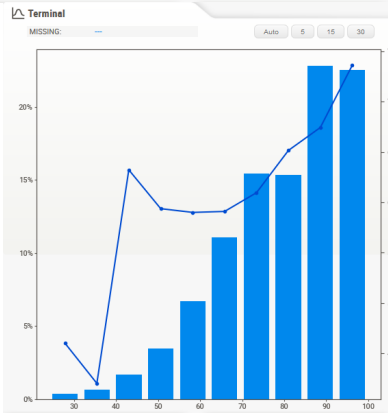
CONTINUED..

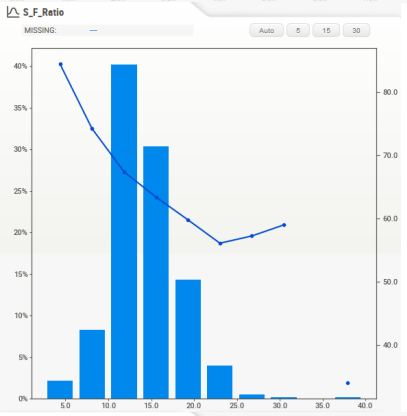
# BIVARIATE ANALYSIS - GRAD_RATE VS REST

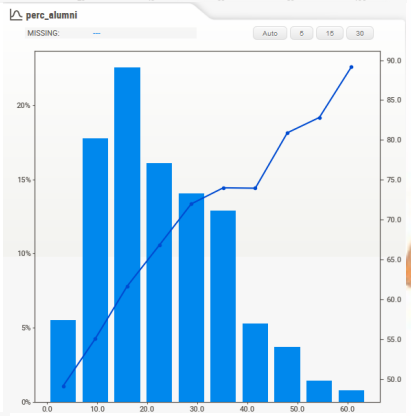


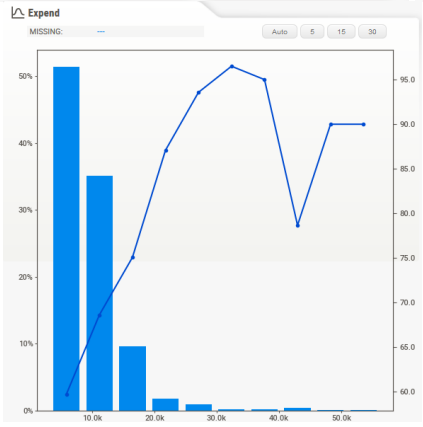



CONTINUED..

## BIVARIATE ANALYSIS - GRAD_RATE VS REST

We have used Python's SweetViz library to generate the above graphs, SweetViz report for Problem 2 can be found in file "SweetViz_Report_Problem2.html".

Following conclusions can be drawn based on the above bivariate analysis.

- More students the college has from Top 10 % of HSC pass outs, higher is the graduation rate.
- More students the college has from Top 25% of HSC pass outs, higher is the graduation rate.
- Higher the number of outstate students for a college, higher is the graduation rate.
    - This in a way tells us, people who have come out of their state specifically to study are more sincere than the people who are studying locally.
- Higher the room and boarding cost, higher is the graduation rate.
    - This in a way tells us, hostels with higher room and boarding cost might be having better facilities enabling students living in them to study more effectively and hence the higher graduation rate for the college.
- Higher the number of PhD faculty, higher is the graduation rate.
- Higher the number of faculty with terminal degree, higher is the graduation rate.
- Higher the Student to Faculty ratio, lower is the graduation rate.
- Higher the Percent of alumni who donate for a college, higher is the graduation rate.

CONTINUED..

## BIVARIATE ANALYSIS - GRAD_RATE VS REST

Apart from the above visual evidence, we also performed ANOVA to confirm the statistical significance of all the above 8 conclusions. Below is the ANOVA output for all the above mentioned conclusions.

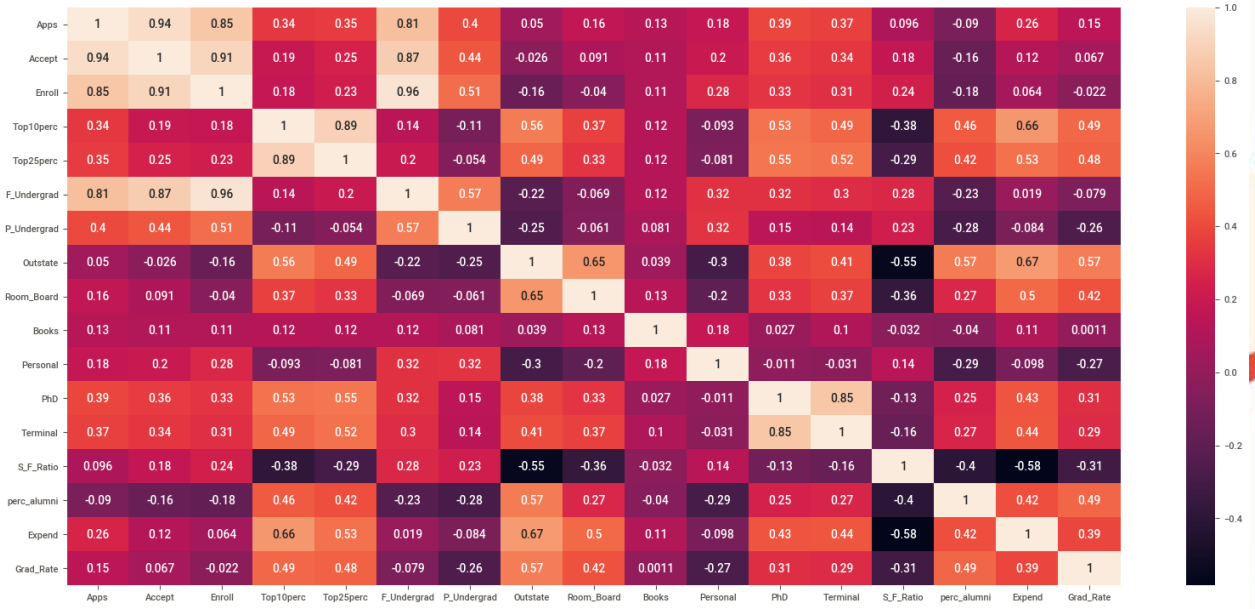
| | p-value | Conclusion |
|---|---|---|
| Top10perc | 2.897974e-49 | Statistically Significant |
| Top25perc | 1.872333e-45 | Statistically Significant |
| Outstate | 1.628927e-68 | Statistically Significant |
| Room_Board | 2.046793e-35 | Statistically Significant |
| PhD | 3.399450e-18 | Statistically Significant |
| Terminal | 1.803054e-16 | Statistically Significant |
| S_F_Ratio | 2.183394e-18 | Statistically Significant |
| perc_alumni | 2.303803e-48 | Statistically Significant |

## BIVARIATE ANALYSIS - CORRELATION

| | 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 | 0.417864 | 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 | 0.417864 | -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 |

CONTINUED..

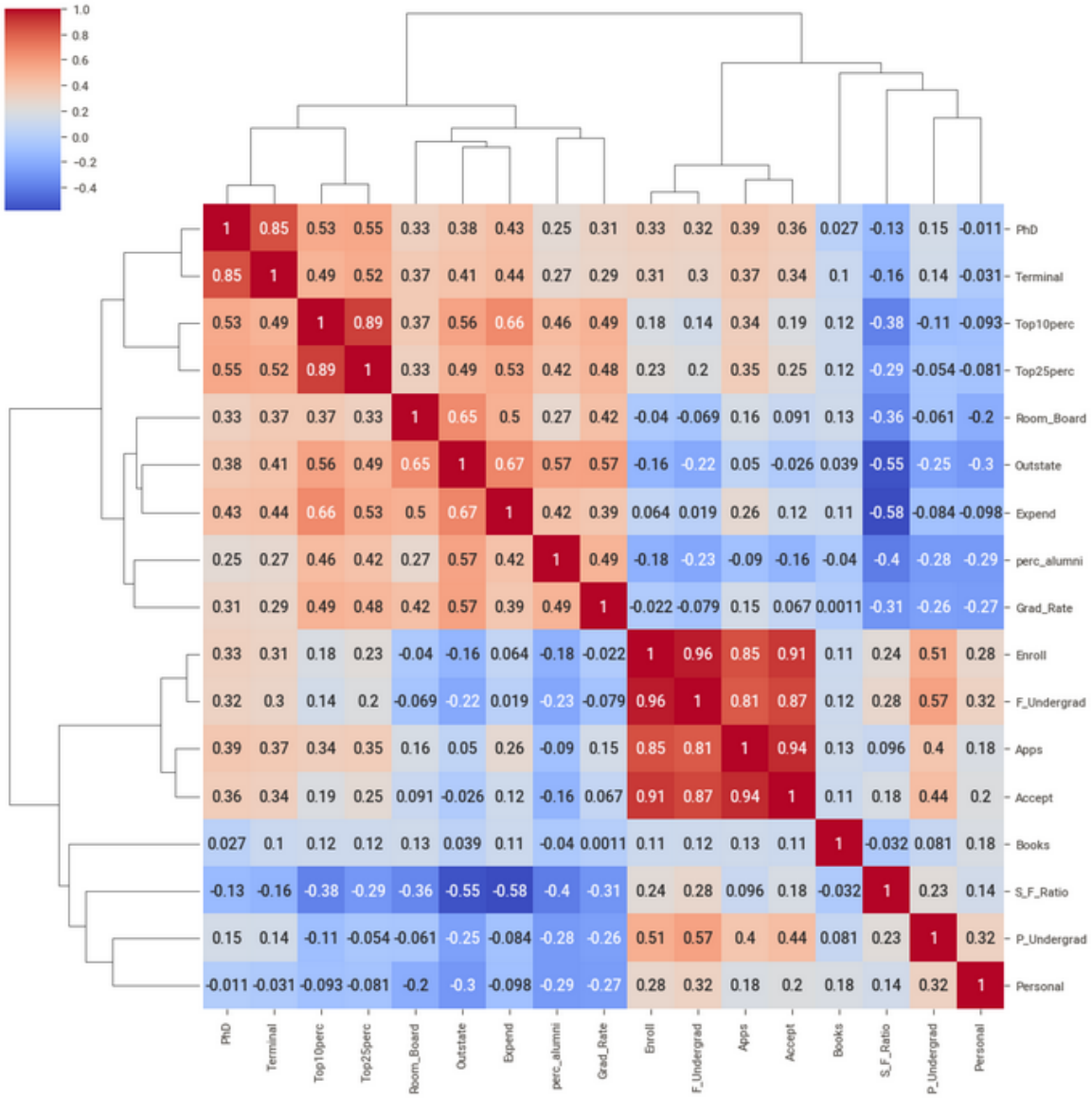
# BIVARIATE ANALYSIS - HEATMAP




- Looking at the correlation matrix and the heat map, we can see some obvious correlations like:
  - F_Undergrad with Apps/Accept/Enrolls.
  - Top10perc with Top25perc
  - Room_Board with Outstate
  - PhD with Terminal
  - perc_alumni with Grad_Rate
  - Outstate with Expend
  - etc.
- Further we are going to create a cluster map to visualize the closely related columns in a better manner.

CONTINUED..

## BIVARIATE ANALYSIS - CLUSTERMAP




- Looking at the above cluster map we can say:-
  - Enroll, F_Undergrad,Apps & Accept variables are most closely related, and are hence shown by the brightest orange box on the map.
  - The other cluster which can be considered to be closely related would consist of Top10perc, Top25perc, Room_Board, Outstate, Expend, perc_alumni & Grad_Rate

# 2.2) SCALE THE VARIABLES AND WRITE THE INFERENCE FOR USING THE TYPE OF SCALING FUNCTION FOR THIS CASE STUDY.

## SCALING - USING STANDARD SCALER

| | 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 | -0.346882 | -0.321205 | -0.063509 | -0.258583 | -0.191827 | -0.168116 | -0.209207 | -0.746356 | -0.964905 | -0.602312 | 1.270045 | -0.163028 | -0.115729 | 1.013776 | -0.867574 | -0.501910 | -0.318252 |
| 1 | -0.210884 | -0.038703 | -0.288584 | -0.655656 | -1.353911 | -0.209788 | 0.244307 | 0.457496 | 1.909208 | 1.215880 | 0.235515 | -2.675646 | -3.378176 | -0.477704 | -0.544572 | 0.166110 | -0.551262 |
| 2 | -0.406866 | -0.376318 | -0.478121 | -0.315307 | -0.292878 | -0.549565 | -0.497090 | 0.201305 | -0.554317 | -0.905344 | -0.259582 | -1.204845 | -0.931341 | -0.300749 | 0.585935 | -0.177290 | -0.667767 |
| 3 | -0.668261 | -0.681682 | -0.692427 | 1.840231 | 1.677612 | -0.658079 | -0.520752 | 0.626633 | 0.996791 | -0.602312 | -0.688173 | 1.185206 | 1.175657 | -1.615274 | 1.151188 | 1.792851 | -0.376504 |
| 4 | -0.726176 | -0.764555 | -0.780735 | -0.655656 | -0.596031 | -0.711924 | 0.009005 | -0.716508 | -0.216723 | 1.518912 | 0.235515 | 0.204672 | -0.523535 | -0.553542 | -1.675079 | 0.241803 | -2.939613 |

Header of the output after using Standard Scaler.

## INFERENCE

- By using the standard scaler, we convert the data into the form where mean for each column is 0 and the standard deviation is 1.
- We scaled the data, since we could see from the original data set that there is a significant difference in the scale of different variables.
- E.g. PhD had a maximum value of 103, while the Expend variable had the maximum value of 56233.
- Also since there are number of outliers in the data, and we would want to preserve that information, we used standard scaler as it tries to preserve the outlier information in the trasnformed data.

## 2.3) COMMENT ON THE COMPARISON BETWEEN COVARIANCE AND THE CORRELATION MATRIX AFTER SCALING.

# COVARIANCE MATRIX OF SCALED DATA

| | 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 |

# CORRELATION MATRIX OF ORIGINAL DATA

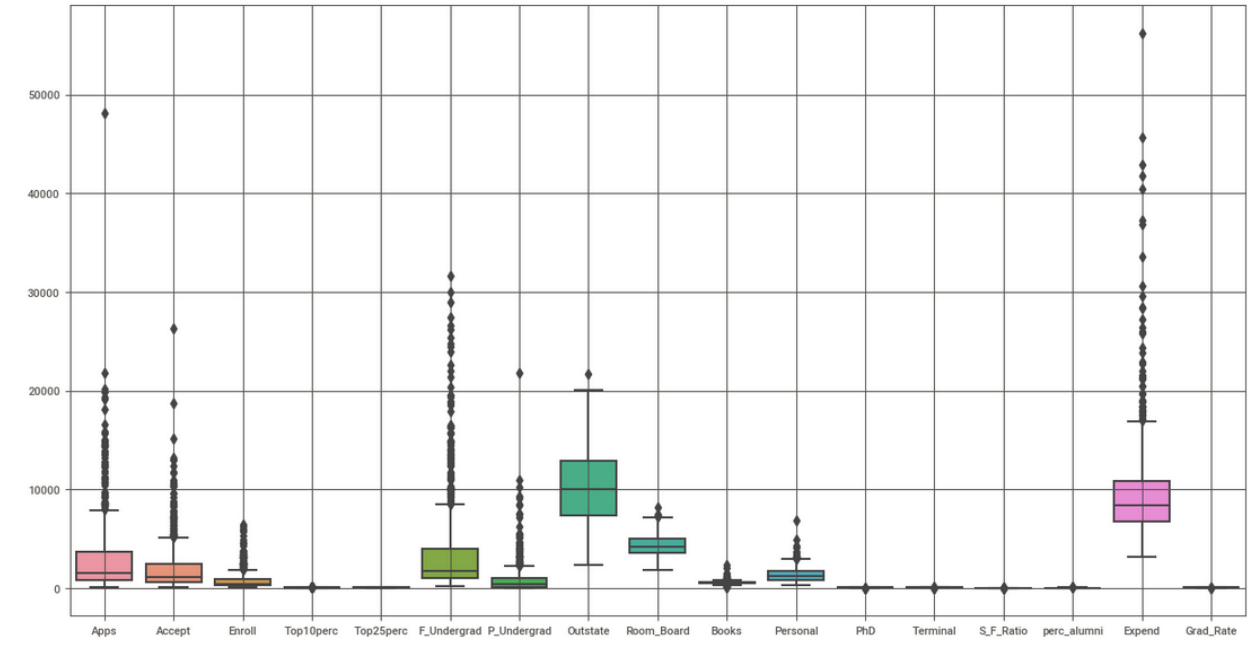
| | Apps | Accept | Enroll | Top10perc | Top25perc | F_Undergrad | P_Undergrad | Outstate | Room_Board | Books | Personal | PhD | Terminal | S_F_Ratio | perc_alumni | Expend | Grad_Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Apps** | 1.000000 | 0.943451 | 0.846822 | 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 | 0.417864 | 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 | 0.115550 | 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 | 0.417864 | -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 |

- Looking at the above two matrices, we can say that the covariance matrix of the standardized data is the same as the correlation matrix of the original data.

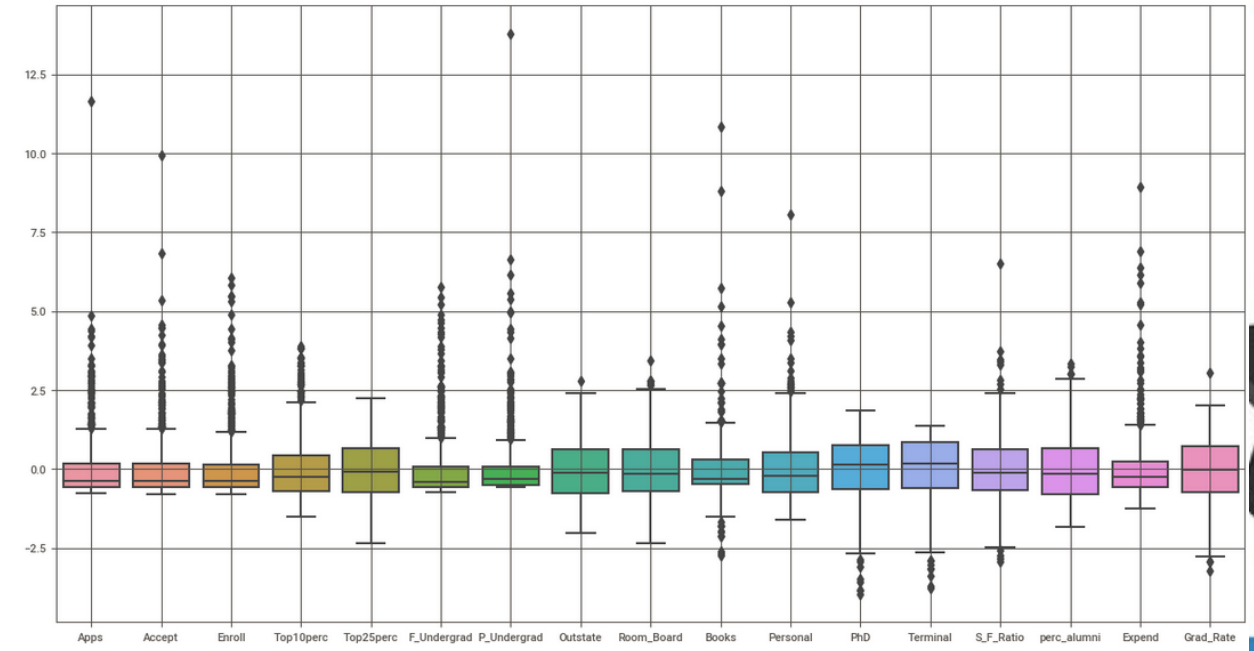
- In Pseudo code, we can say:
  - cov_matrix(scaled_data) = orginal_data.corr()

## 2.4) CHECK THE DATASET FOR OUTLIERS BEFORE AND AFTER SCALING. DRAW YOUR INFERENCES FROM THIS EXERCISE.

## OUTLIERS BEFORE SCALING




## OUTLIERS AFTER SCALING



CONTINUED..

## INFERENCES

- We can clearly see the outlier information is largely preserved after we had scaled the data using standard scaler.
- Also visualizing the outlier is much easier once the data has been standardized, compared to before.

| | Percentage of Outliers Before Scaling |
|---|---|
| F_Undergrad | 12.48 |
| Enroll | 10.17 |
| Accept | 9.40 |
| Apps | 9.01 |
| P_Undergrad | 8.62 |
| Expend | 6.18 |
| Books | 5.92 |
| Top10perc | 5.02 |
| Personal | 2.57 |
| S_F_Ratio | 1.54 |
| PhD | 1.03 |
| Terminal | 1.03 |
| Room_Board | 0.90 |
| perc_alumni | 0.64 |
| Grad_Rate | 0.51 |
| Outstate | 0.13 |
| Top25perc | 0.00 |

| | Percentage of Outliers After Scaling |
|---|---|
| F_Undergrad | 12.48 |
| Enroll | 10.17 |
| Accept | 9.40 |
| Apps | 9.01 |
| P_Undergrad | 8.62 |
| Books | 6.18 |
| Expend | 6.18 |
| Top10perc | 5.02 |
| Personal | 2.57 |
| S_F_Ratio | 1.54 |
| PhD | 1.03 |
| Terminal | 1.03 |
| Room_Board | 0.90 |
| perc_alumni | 0.64 |
| Grad_Rate | 0.51 |
| Outstate | 0.13 |
| Top25perc | 0.00 |

- From the above analysis we can confirm the number of outliers remain more or less the same after scaling using standard scaler.
- Also since we have quite a large number of outliers for multiple columns, we chose not to do outlier treatment before proceeding for PCA as that might tamper with the essence of the data.

## 2.5) BUILD THE COVARIANCE MATRIX AND CALCULATE THE EIGENVALUES AND THE EIGENVECTOR.

# COVARIANCE MATRIX

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554(
    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  1.00128866 -0.01094989
   -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
 [ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
    0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
    0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
 [ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
    0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
    1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
 [ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
    0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
   -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
[-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
   -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
    0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
 [ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
   -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
    0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
 [ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
   -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
    0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

CONTINUED..

# EIGEN VECTOR

```
Eigen Vectors
%s [[-2.48765602e-01 -2.07601502e-01 -1.76303592e-01 -3.54273947e-
  -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  3.02385408e-01  2.22532003e-01
   5.60919470e-01 -1.27288825e-01 -2.22311021e-01  1.40166326e-01
   2.04719730e-01 -7.93882496e-02 -2.16297411e-01  7.59581203e-02
  -1.09267913e-01]
 [ 1.62374420e-02 -7.53468452e-03  4.25579803e-02  5.26927980e-02
  -3.30915896e-02  4.34542349e-02  1.91198583e-01  3.00003910e-02
  -1.62755446e-01 -6.41054950e-01  3.31398003e-01 -9.12555212e-02
  -1.54927646e-01 -4.87045875e-01  4.73400144e-02  2.98118619e-01
  -2.16163313e-01]
 [ 4.24863486e-02  1.29497196e-02  2.76928937e-02  1.61332069e-01
   1.18485556e-01  2.50763629e-02 -6.10423460e-02 -1.08528966e-01
  -2.09744235e-01  1.49692034e-01 -6.33790064e-01  1.09641298e-03
   2.84770105e-02 -2.19259358e-01 -2.43321156e-01  2.26584481e-01
  -5.59943937e-01]
 [ 1.03090398e-01  5.62709623e-02 -5.86623552e-02  1.22678028e-01
   1.02491967e-01 -7.88896442e-02 -5.70783816e-01 -9.84599754e-03
   2.21453442e-01 -2.13293009e-01  2.32660840e-01  7.70400002e-02
   1.21613297e-02  8.36048735e-02 -6.78523654e-01  5.41593771e-02
   5.33553891e-03]
 [ 9.02270802e-02  1.77864814e-01  1.28560713e-01 -3.41099863e-01
  -4.03711989e-01  5.94419181e-02 -5.60672902e-01  4.57332880e-03
  -2.75022548e-01  1.33663353e-01  9.44688900e-02  1.85181525e-01
   2.54938198e-01 -2.74544380e-01  2.55334907e-01  4.91388809e-02
  -4.19043052e-02]
 [-5.25098025e-02 -4.11400844e-02 -3.44879147e-02 -6.40257785e-02
  -1.45492289e-02 -2.08471834e-02  2.23105808e-01 -1.86675363e-01
  -2.98324237e-01  8.20292186e-02 -1.36027616e-01  1.23452200e-01
   8.85784627e-02 -4.72045249e-01 -4.22999706e-01 -1.32286331e-01
   5.90271067e-01]
[ 3.58970400e-01 -5.43427250e-01  6.09651110e-01 -1.44986329e-01
  8.03478445e-02 -4.14705279e-01  9.01788964e-03  5.08995918e-02
  1.14639620e-03  7.72631963e-04 -1.11433396e-03  1.38133366e-02
  6.20932749e-03 -2.22215182e-03 -1.91869743e-02 -3.53098218e-02
 -1.30710024e-02]
[-4.59139498e-01  5.18568789e-01  4.04318439e-01  1.48738723e-01
 -5.18683400e-02 -5.60363054e-01  5.27313042e-02 -1.01594830e-01
  2.59293381e-02 -2.88282896e-03  1.28904022e-02 -2.98075465e-02
  2.70759809e-02  2.12476294e-02 -3.33406243e-03  4.38803230e-02
  5.00844705e-03]
[ 4.30462074e-02 -5.84055850e-02 -6.93988831e-02 -8.10481404e-03
 -2.73128469e-01 -8.11578181e-02  1.00693324e-01  1.43220673e-01
 -3.59321731e-01  3.19400370e-02 -1.85784733e-02  4.03723253e-02
 -5.89734026e-02  4.45000727e-01 -1.30727978e-01  6.92088870e-01
  2.19839000e-01]
[-1.33405806e-01  1.45497511e-01 -2.95896092e-02 -6.97722522e-01
  6.17274818e-01 -9.91640992e-03 -2.09515982e-02 -3.83544794e-02
 -3.40197083e-03  9.43887925e-03  3.09001353e-03  1.12055599e-01
 -1.58909651e-01  2.08991284e-02  8.41789410e-03  2.27742017e-01
  3.39433604e-03]
[ 8.06328039e-02  3.34674281e-02 -8.56967180e-02 -1.07828189e-01
  1.51742110e-01 -5.63728817e-02  1.92857500e-02 -3.40115407e-02
 -5.84289756e-02 -6.68494643e-02  2.75286207e-02 -6.91126145e-01
  6.71008607e-01  4.13740967e-02 -2.71542091e-02  7.31225166e-02
  3.64767385e-02]
[-5.95830975e-01 -2.92642398e-01  4.44638207e-01 -1.02303616e-01
 -2.18838802e-02  5.23622267e-01 -1.25997650e-01  1.41856014e-01
  6.97485854e-02 -1.14379958e-02 -3.94547417e-02 -1.27696382e-01
  5.83134662e-02  1.77152700e-02 -1.04088088e-01  9.37464497e-02
  6.91969778e-02]
[ 2.40709086e-02 -1.45102446e-01  1.11431545e-02  3.85543001e-02
 -8.93515563e-02  5.61767721e-02 -6.35360730e-02 -8.23443779e-01
  3.54559731e-01 -2.81593679e-02 -3.92640266e-02  2.32224316e-02
  1.64850420e-02 -1.10262122e-02  1.82660654e-01  3.25982295e-01
  1.22106697e-01]]
```

CONTINUED..

# EIGEN VALUES

```
 Eigen Values
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
 0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
 0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

# 2.6) WRITE THE EXPLICIT FORM OF THE FIRST PC (IN TERMS OF EIGEN VECTORS).

## EXPLICIT FORM OF FIRST PC

| | 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 | 0.248766 | 0.207602 | 0.176304 | 0.354274 | 0.344001 | 0.154641 | 0.026443 | 0.294736 | 0.249030 | 0.064758 | -0.042529 | 0.318313 | 0.317056 | -0.176958 | 0.205082 | 0.318909 | 0.252316 |

- **First Principal Component - PC1**
  - 0.248766 * Apps + 0.207602 * Accept + 0.176304 * Enroll + 0.354274 * Top10perc + 0.344001 * Top25perc + 0.154641 * F_Undergrad + 0.026443 * P_Undergrad + 0.294736 * Outstate + 0.249030 * Room_Board + 0.064758 * Books + 0.042529 * Personal + 0.318313 * PhD + 0.317056 * Terminal - 0.176958 * S_F_Ratio + 0.205082 * perc_alumni + 0.318909 * Expend + 0.252316 * Grad_Rate

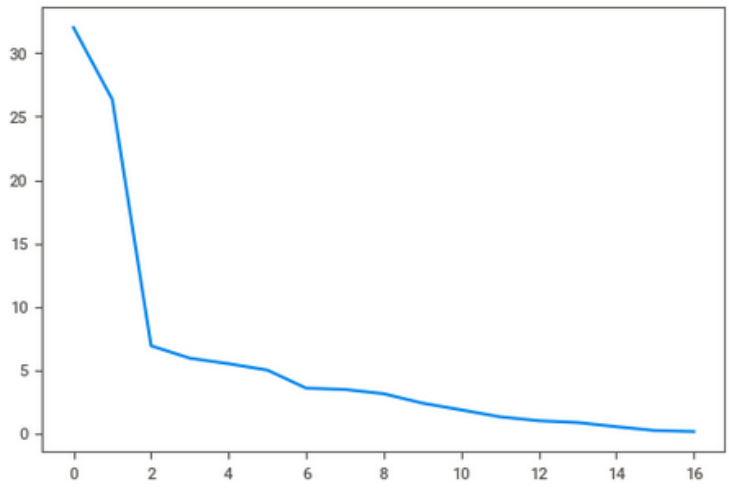
2.7) DISCUSS THE CUMULATIVE VALUES OF THE EIGENVALUES. HOW DOES IT HELP YOU TO DECIDE ON THE OPTIMUM NUMBER OF PRINCIPAL COMPONENTS? WHAT DO THE EIGENVECTORS INDICATE? PERFORM PCA AND EXPORT THE DATA OF THE PRINCIPAL COMPONENT SCORES INTO A DATA FRAME.

# CUMULATIVE VARIANCE EXPLAINED

```
Cumulative Variance Explained [ 32.0206282   58.36084263  65.26175919  71.18474841  76.67315352
  81.65785448  85.21672597  88.67034731  91.78758099  94.16277251
  96.00419883  97.30024023  98.28599436  99.13183669  99.64896227
  99.86471628 100.        ]
```

Cumulative variance explains the amount of variance that is being captured by each of the principal component added together in a cumulative fashion. i.e. first element in the array gives the total variance explained by PC1, second element gives the total variance explained by PC1 and PC2 combined, so on and so forth.

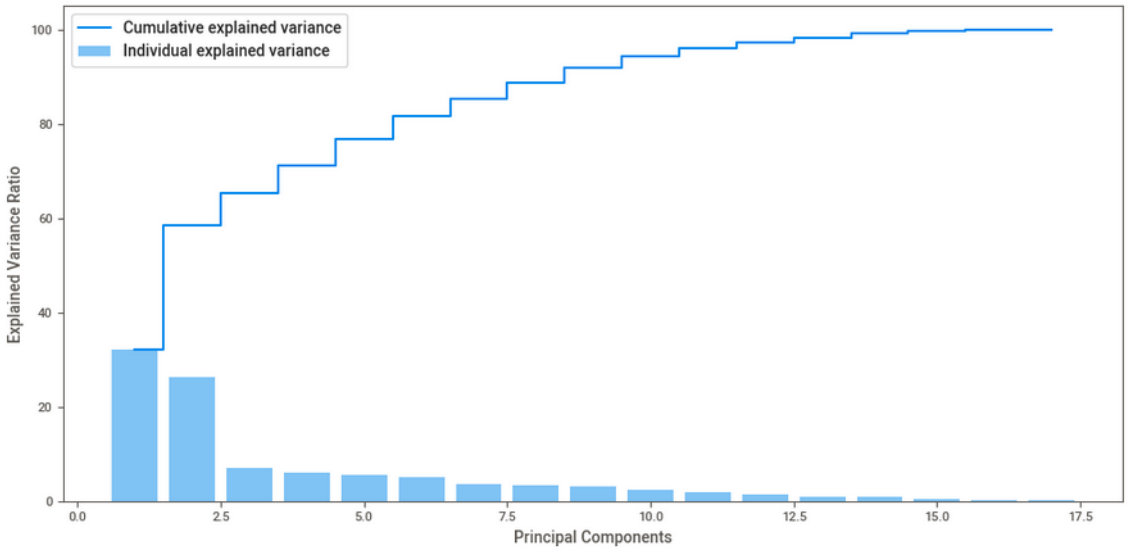
```
Variance Explained [32.02062819886915, 26.340214436112465, 6.900916554222497, 5.922989222926289, 5.488405110358481,
                    4.98470095455745, 3.55887149174665, 3.4536213369992637, 3.1172336798217195, 2.3751915258937997,
                    1.841426, 3209386879, 1.2960414001235345, 0.9857541228001161, 0.8458423350830022, 0.517125583373194,
                    0.2157540100727585, 0.13528371610095183]
```




- Plotting a scree plot from the values found in the variance explained array.
- From the graph we can see, most of the variance is being captured by first 2 principal components, after which the variance explained drops sharply.

CONTINUED..

# EXPLAINED VARIANCE RATIO




- The bars in the above plot shows the individual variance explained by each principal component.
- The line above the bars, shows the cumulative variance explained as we move along the x-axis, which indicated the different PCs.
- Using the above graph and the concept of cumulative variance, we can find the number of principal component we need to use.
- The amount of variance that needs to be explained by the PCs will vary from experiment to experiment, depending on the requirements.
- For e.g. let us assume the required variance to be captured is 90%, now looking at the graph above, we can see if we take into account the first 9 PCs we can capture around 91% variance of the data.
- In the process we also managed to reduce the total number of dimensions from 17 to 9.
- This is how we can use the concept of cumulative variance explained to decide on the number of Principal components that we need to use.

CONTINUED..

# WHAT DO EIGEN VECTORS INDICATE

The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude.

In other words, the eigenvalues explain the variance of the data along the new feature axes.

# PRINCIPAL COMPONENTS

| | 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 | 0.248766 | 0.207602 | 0.176304 | 0.354274 | 0.344001 | 0.154641 | 0.026443 | 0.294736 | 0.249030 | 0.064758 | -0.042529 | 0.318313 | 0.317056 | -0.176958 | 0.205082 | 0.318909 | 0.252316 |
| 1 | 0.331598 | 0.372117 | 0.403724 | -0.082412 | -0.044779 | 0.417674 | 0.315088 | -0.249644 | -0.137809 | 0.056342 | 0.219929 | 0.058311 | 0.046429 | 0.246665 | -0.246595 | -0.131690 | -0.169241 |
| 2 | -0.063092 | -0.101249 | -0.082986 | 0.035056 | -0.024148 | -0.061393 | 0.139682 | 0.046599 | 0.148967 | 0.677412 | 0.499721 | -0.127028 | -0.066038 | -0.289848 | -0.146989 | 0.226744 | -0.208065 |
| 3 | 0.281311 | 0.267817 | 0.161827 | -0.051547 | -0.109767 | 0.100412 | -0.158558 | 0.131291 | 0.184996 | 0.087089 | -0.230711 | -0.534725 | -0.519443 | -0.161189 | 0.017314 | 0.079273 | 0.269129 |
| 4 | 0.005741 | 0.055786 | -0.055694 | -0.395434 | -0.426534 | -0.043454 | 0.302385 | 0.222532 | 0.560919 | -0.127289 | -0.222311 | 0.140166 | 0.204720 | -0.079388 | -0.216297 | 0.075958 | -0.109268 |
| 5 | -0.016237 | 0.007535 | -0.042558 | -0.052693 | 0.033092 | -0.043454 | -0.191199 | -0.030000 | 0.162755 | 0.641055 | -0.331398 | 0.091256 | 0.154928 | 0.487046 | -0.047340 | -0.298119 | 0.216163 |
| 6 | -0.042486 | -0.012950 | -0.027693 | -0.161332 | -0.118486 | -0.025076 | 0.061042 | 0.108529 | 0.209744 | -0.149692 | 0.633790 | -0.001096 | -0.028477 | 0.219259 | 0.243321 | -0.226584 | 0.559944 |
| 7 | -0.103090 | -0.056271 | 0.058662 | -0.122678 | -0.102492 | 0.078890 | 0.570784 | 0.009846 | -0.221453 | 0.213293 | -0.232661 | -0.077040 | -0.012161 | -0.083605 | 0.678524 | -0.054159 | -0.005336 |
| 8 | -0.090227 | -0.177865 | -0.128561 | 0.341100 | 0.403712 | -0.059442 | 0.560673 | -0.004573 | 0.275023 | -0.133663 | -0.094469 | -0.185182 | -0.254938 | 0.274544 | -0.255335 | -0.049139 | 0.041904 |
| 9 | 0.052510 | 0.041140 | 0.034488 | 0.064026 | 0.014549 | 0.020847 | -0.223106 | 0.186675 | 0.298324 | -0.082029 | 0.136028 | -0.123452 | -0.088578 | 0.472045 | 0.423000 | 0.132286 | -0.590271 |
| 10 | 0.043046 | -0.058406 | -0.069399 | -0.008105 | -0.273128 | -0.081158 | 0.100693 | 0.143221 | -0.359322 | 0.031940 | -0.018578 | 0.040372 | -0.058973 | 0.445001 | -0.130728 | 0.692089 | 0.219839 |
| 11 | 0.024071 | -0.145102 | 0.011143 | 0.038554 | -0.089352 | 0.056177 | -0.063536 | -0.823444 | 0.354560 | -0.028159 | -0.039264 | 0.023222 | 0.016485 | -0.011026 | 0.182661 | 0.325982 | 0.122107 |
| 12 | 0.595831 | 0.292642 | -0.444638 | 0.001023 | 0.021884 | -0.523622 | 0.125998 | -0.141856 | -0.069749 | 0.011438 | 0.039455 | 0.127696 | -0.058313 | -0.017715 | 0.104088 | -0.093746 | -0.069197 |
| 13 | 0.080633 | 0.033467 | -0.085697 | -0.107828 | 0.151742 | -0.056373 | 0.019286 | -0.034012 | -0.058429 | -0.066849 | 0.027529 | -0.691126 | 0.671009 | 0.041374 | -0.027154 | 0.073123 | 0.036477 |
| 14 | 0.133406 | -0.145498 | 0.029590 | 0.697723 | -0.617275 | 0.009916 | 0.020952 | 0.038354 | 0.003402 | -0.009439 | -0.003090 | -0.112056 | 0.158910 | -0.020899 | -0.008418 | -0.227742 | -0.003394 |
| 15 | 0.459139 | -0.518569 | -0.404318 | -0.148739 | 0.051868 | 0.560363 | -0.052731 | 0.101595 | -0.025929 | 0.002883 | -0.012890 | 0.029808 | -0.027076 | -0.021248 | 0.003334 | -0.043880 | -0.005008 |
| 16 | 0.358970 | -0.543427 | 0.609651 | -0.144986 | 0.080348 | -0.414705 | 0.009018 | 0.050900 | 0.001146 | 0.000773 | -0.001114 | 0.013813 | 0.006209 | -0.002222 | -0.019187 | -0.035310 | -0.013071 |

DataFrame showing all the principal components generated for this case.

We might choose to use only top 9 PCs since that is already capturing 90% of the variance in the data, assuming 90% of variance is more than sufficient in this case.

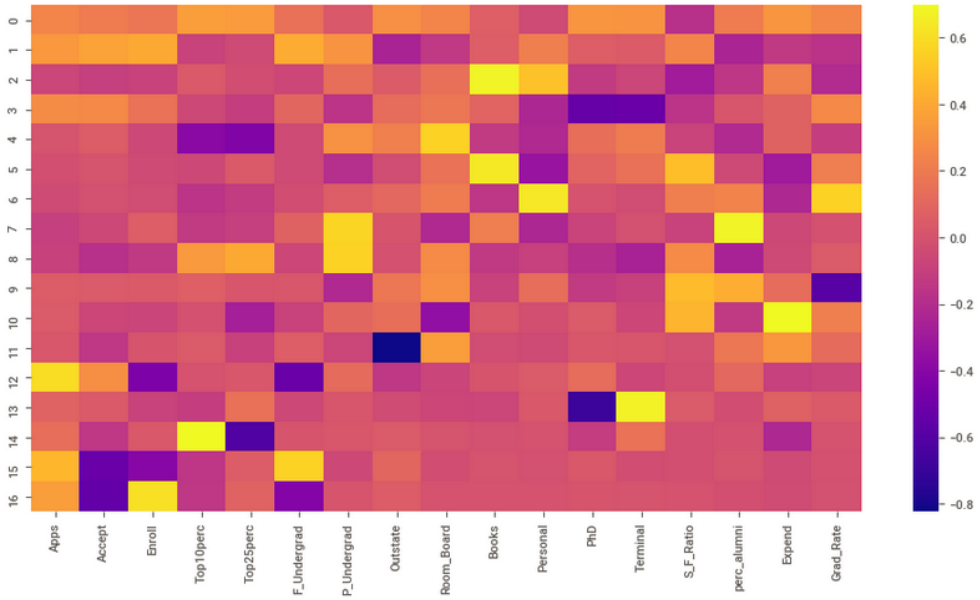
To increase the variance captured, we can choose to increase the number of PCs.

CONTINUED..

# FACTOR PLOT

A factor plot shows the relationship between different PCs and the various columns, we can see which PC captures more variance for which particular columns.

Factor plot for different Principal Components. We can choose a particular PC based on the variable in which we are interested. For e.g. if we are interested in say column Books, then we would use PC3 as it captures the maximum variance for this column.



2.8) MENTION THE BUSINESS IMPLICATION OF USING THE PRINCIPAL COMPONENT ANALYSIS FOR THIS CASE STUDY.

## BUSINESS IMPLICATIONS

- 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.
- Depending on what variable the business is interested in, we can actually make use of different principal components.
- For e.g. if the college is interested to capture more information about how cost of book per student impacts the graduation rate we can make use of PC3 which captures more variance for the variable Books.
- Similarly we can do this for other variables as well.
- Using principal component analysis, we can also see the influential feature for decision making. For e.g. we know the first PC captures the most variance in the data, and hence the columns for which we have the highest co-efficient in PC1, will become the most influential feature.
- We also did bivariate analysis on the data, which revealed various correlations like, Top10perc with Graduation Rate, impact of the number of faculty having PhD on the graduation rate etc. These are also useful insights for the business for decision making.

# THANK YOU!