EFFECT OF A & B COMPOUND ON HAY FEVER

Null & Alternate Hypothesis for Compund A

- μA1-Mean hours of Relief for Treatment A at level 1
  μA2-Mean hours of Relief for Treatment A at level 2
  μA3-Mean hours of Relief for Treatment A at level 3
- Null and Alternate Hypothesis for conducting Treatment A
  H0-μA1=μA2=μA3
  Ha-(μA1≠μA2≠μA3) or (μA1≠μA2=μA3) or (μA1=μA2≠μA3) or (μA1=μA3≠μA2)
  At least one pair of means not equal

Null & Alternate Hypothesis for Compound B

- μB1-Mean hours of Relief for Treatment B at level 1
  μB2-Mean hours of Relief for Treatment B at level 2
  μB3-Mean hours of Relief for Treatment B at level 3
- Null and Alternate Hypothesis for conducting Treatment B
  H0-μB1=μB2=μB3
  Ha-(μB1≠μB2≠μA3) or (μB1≠μB2=μB3) or (μB1=μB2≠μB3) or (μB1=μB3≠μB2)
  At least one pair of means not equal

One-way Anova of Comp A w.r.t Relief

- By looking at below results , we could reject the Null Hypothesis at 5 % significance level. In other words we could say when we take at least one pair , the relief provided by one level of compound A is different than the other level of Compound A.

	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

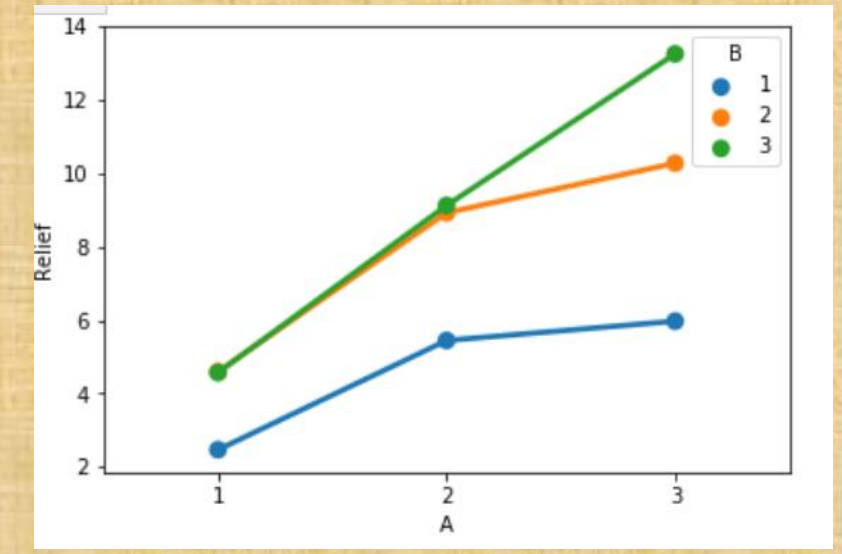
One-way Anova of Comp B w.r.t Relief

- By looking at below results , we could reject the Null Hypothesis at 5 % significance level. In other words we could say when we take at least one pair , the relief provided by one level of compound B is different than the other level of Compound B:

	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

Interaction between 2 Treatments A & B

- From below we could see when compound A of level 3 is combined with compound B of level 3 it can be observed that the combination provides more relief than when level 1 and 2 of compounds A and B interacts. While minimum relief is provided when level 1 of both comp A and B are combined:



2 Way ANOVA based on Comp A & B

Without Interaction

- The compound A and compound B when varied at different levels individually relief hours also changed so we could reject the null hypothesis(at 5% significance level):

	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

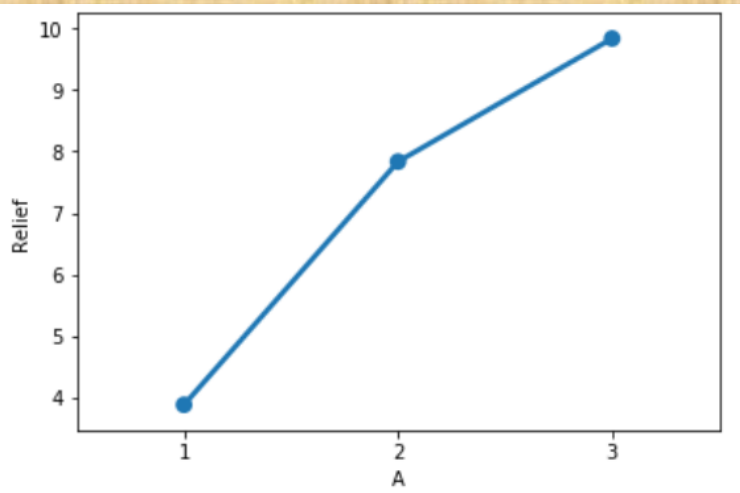
With Interaction:

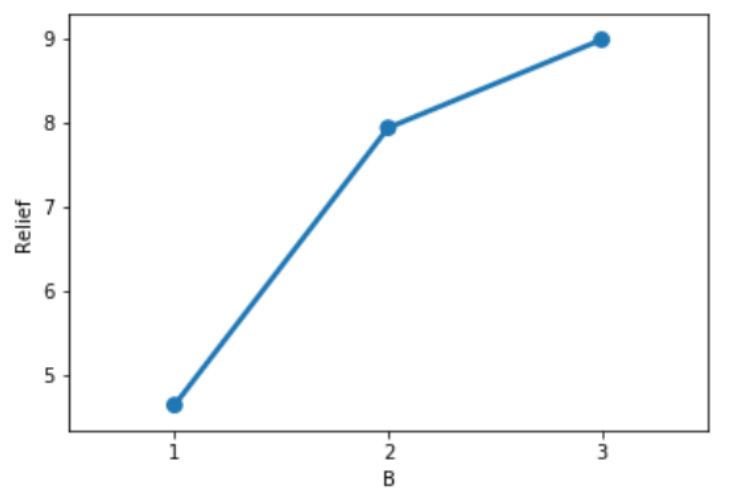
- From below results we could say that when compounds A and B are allowed to interact with each other varied level there is a significant change in relief hours when each combination is considered so we could reject the null hypothesis(at 5% significance level)

	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

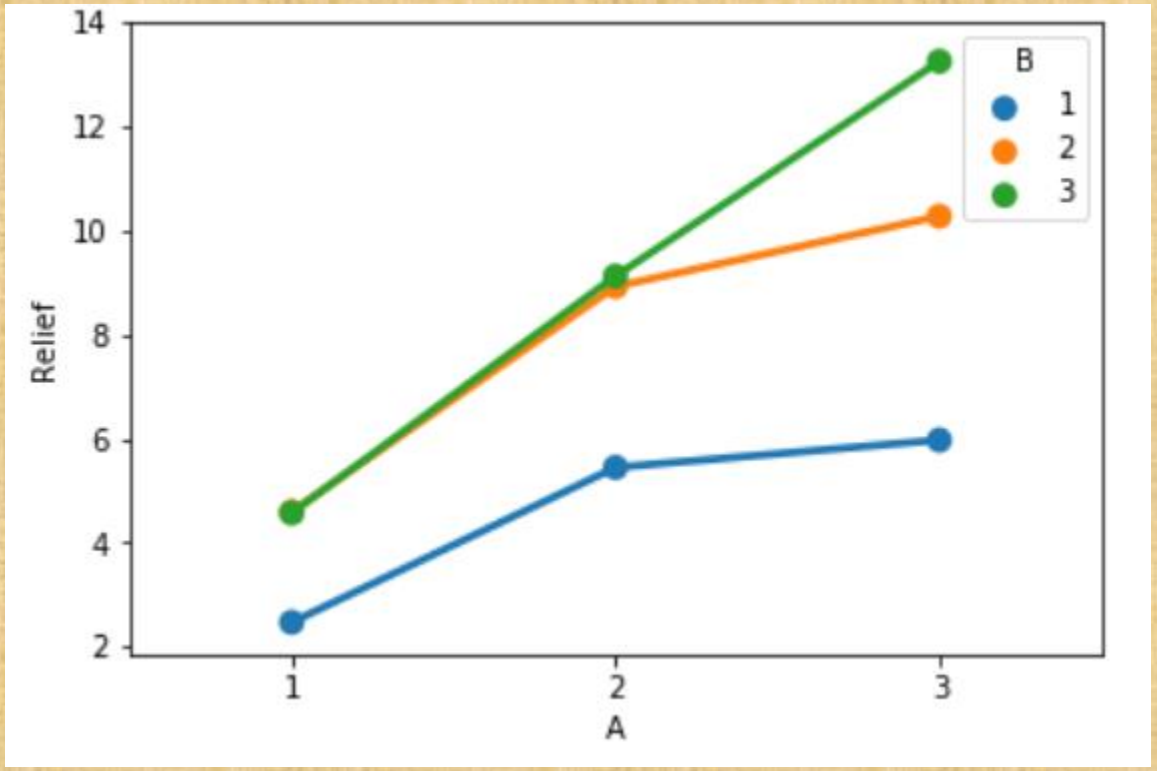
BUSINESS IMPLICATIONS/CONCLUSION

- When treated with compound A and compound B individually we could see there is a increase in relief hours as we increase the levels for a compound:





- When compound A is made to interact with compound B we could see there is a significant increase in relief when level 3 of the compounds are made to interact, there is no or less increase when level 1 of the compounds A and B are made to interact:



THE END

PCA ON PARAMETERS OF VARIOUS COLLEGES

Exploratory Data Analysis

- We are able to see there no missing values in the dataset,no junk values present and also no duplicates present. Dataset is a ‘clean Dataset’:

```
Names          777 non-null object
Apps           777 non-null int64
Accept         777 non-null int64
Enroll         777 non-null int64
Top10perc      777 non-null int64
Top25perc      777 non-null int64
F.Undergrad    777 non-null int64
P.Undergrad    777 non-null int64
Outstate       777 non-null int64
Room.Board     777 non-null int64
Books          777 non-null int64
Personal       777 non-null int64
PhD            777 non-null int64
Terminal       777 non-null int64
S.F.Ratio      777 non-null float64
perc.alumni    777 non-null int64
Expend         777 non-null int64
Grad.Rate      777 non-null int64
```

- From below 5 point summary of the dataset we could see the no:of Applications, Accepted, no:of Full time Undergraduates, no:of Part time Undergraduates and Expenditure on Students shows maximum variation

```
df.median()
```

Apps	1558.0
Accept	1110.0
Enroll	434.0
Top10perc	23.0
Top25perc	54.0
F.Undergrad	1707.0
P.Undergrad	353.0
Outstate	9990.0
Room.Board	4200.0
Books	500.0
Personal	1200.0
PhD	75.0
Terminal	82.0
S.F.Ratio	13.6
perc.alumni	21.0
Expend	8377.0
Grad.Rate	65.0

```
#IQR
df.quantile(0.75)-df.quantile(0.25)
```

Apps	2848.0
Accept	1820.0
Enroll	660.0
Top10perc	20.0
Top25perc	28.0
F.Undergrad	3013.0
P.Undergrad	872.0
Outstate	5605.0
Room.Board	1453.0
Books	130.0
Personal	850.0
PhD	23.0
Terminal	21.0
S.F.Ratio	5.0
perc.alumni	18.0
Expend	4079.0
Grad.Rate	25.0

dtype: float64

```
#Range
df.max(numeric_only=True)-df.min(numeric_only=True)
```

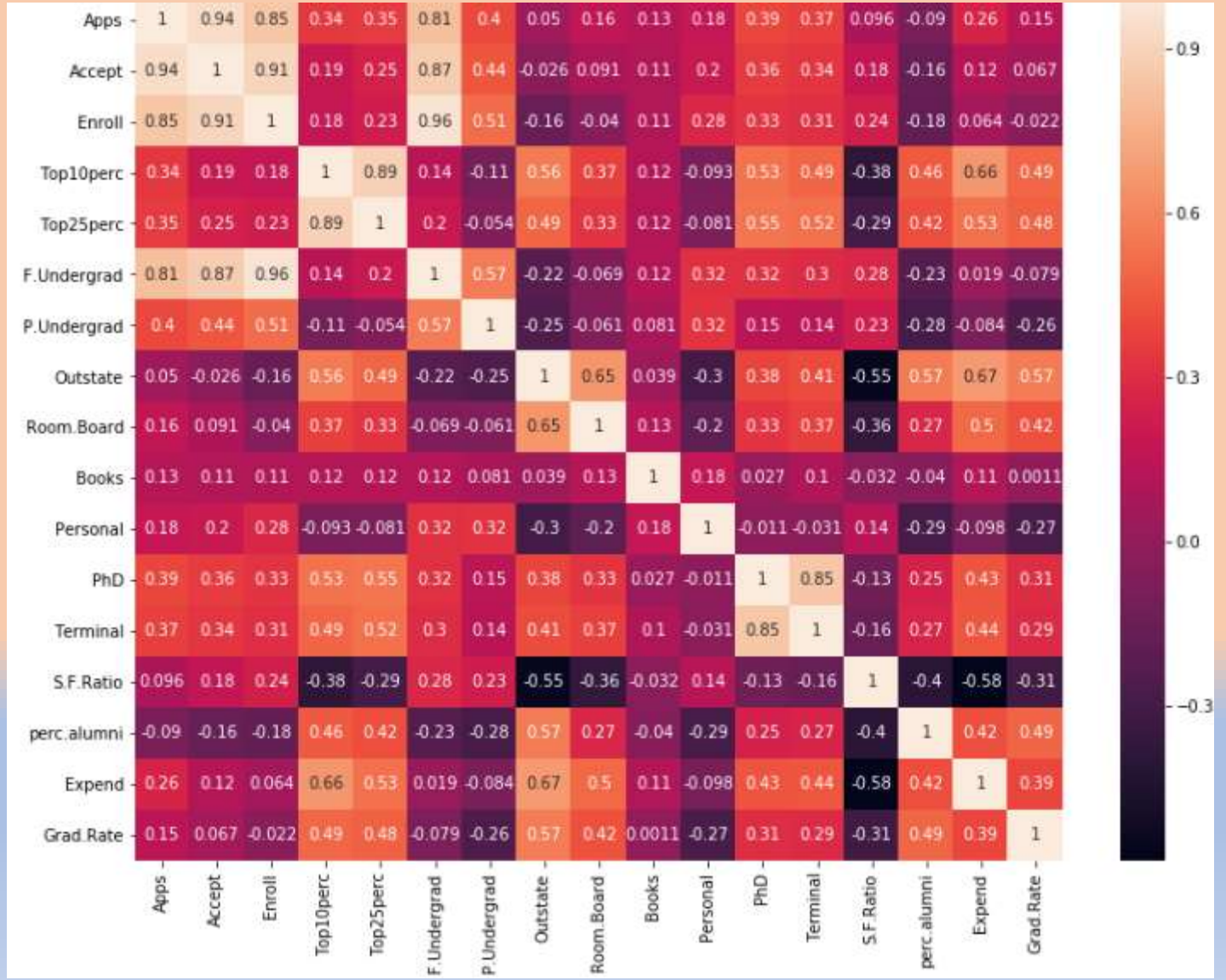
Apps	48013.0
Accept	26258.0
Enroll	6357.0
Top10perc	95.0
Top25perc	91.0
F.Undergrad	31504.0
P.Undergrad	21835.0
Outstate	19360.0
Room.Board	6344.0
Books	2244.0
Personal	6550.0
PhD	95.0
Terminal	76.0
S.F.Ratio	37.3
perc.alumni	64.0
Expend	53047.0
Grad.Rate	108.0

dtype: float64

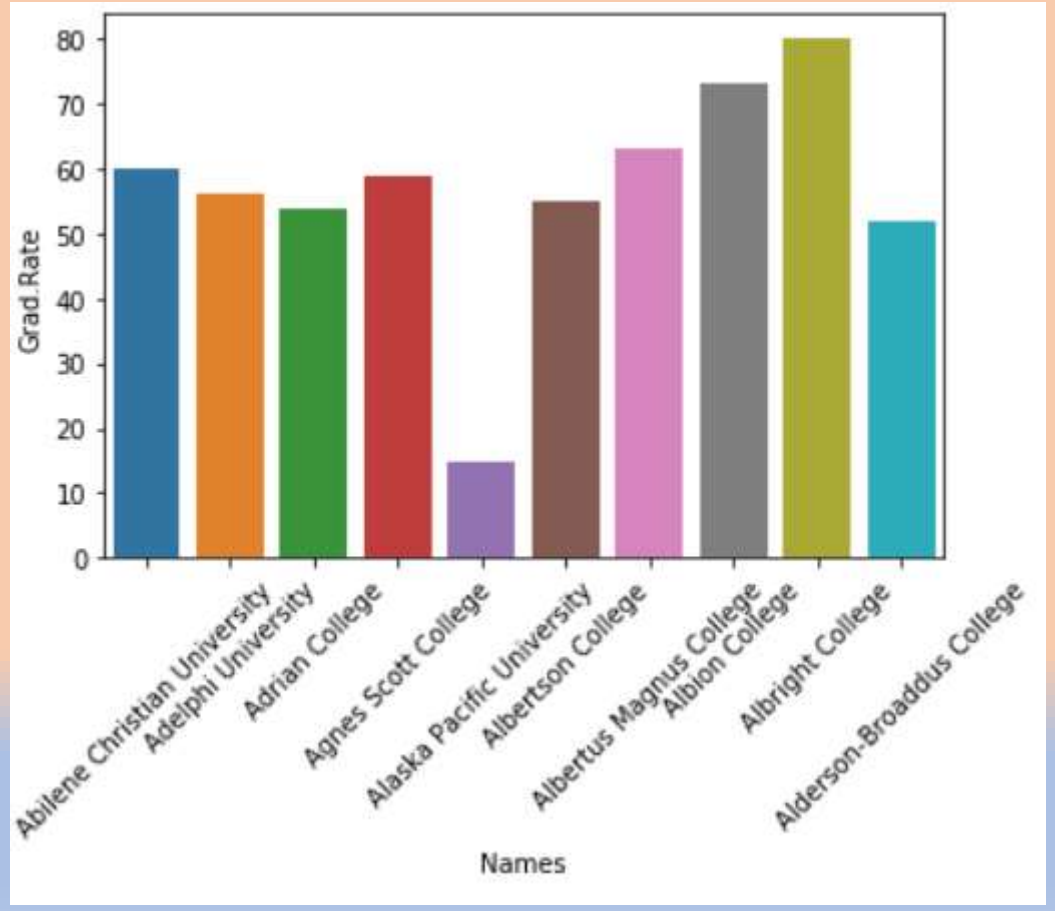
- Except Top25Perc we could see all other variables has outliers as we could see from below where Full time Undergraduates has the most number of outliers:

Apps	70
Accept	73
Enroll	79
Top10perc	39
Top25perc	0
F.Undergrad	97
P.Undergrad	67
Outstate	1
Room.Board	7
Books	46
Personal	20
PhD	8
Terminal	8
S.F.Ratio	12
perc.alumni	5
Expend	48
Grad.Rate	4

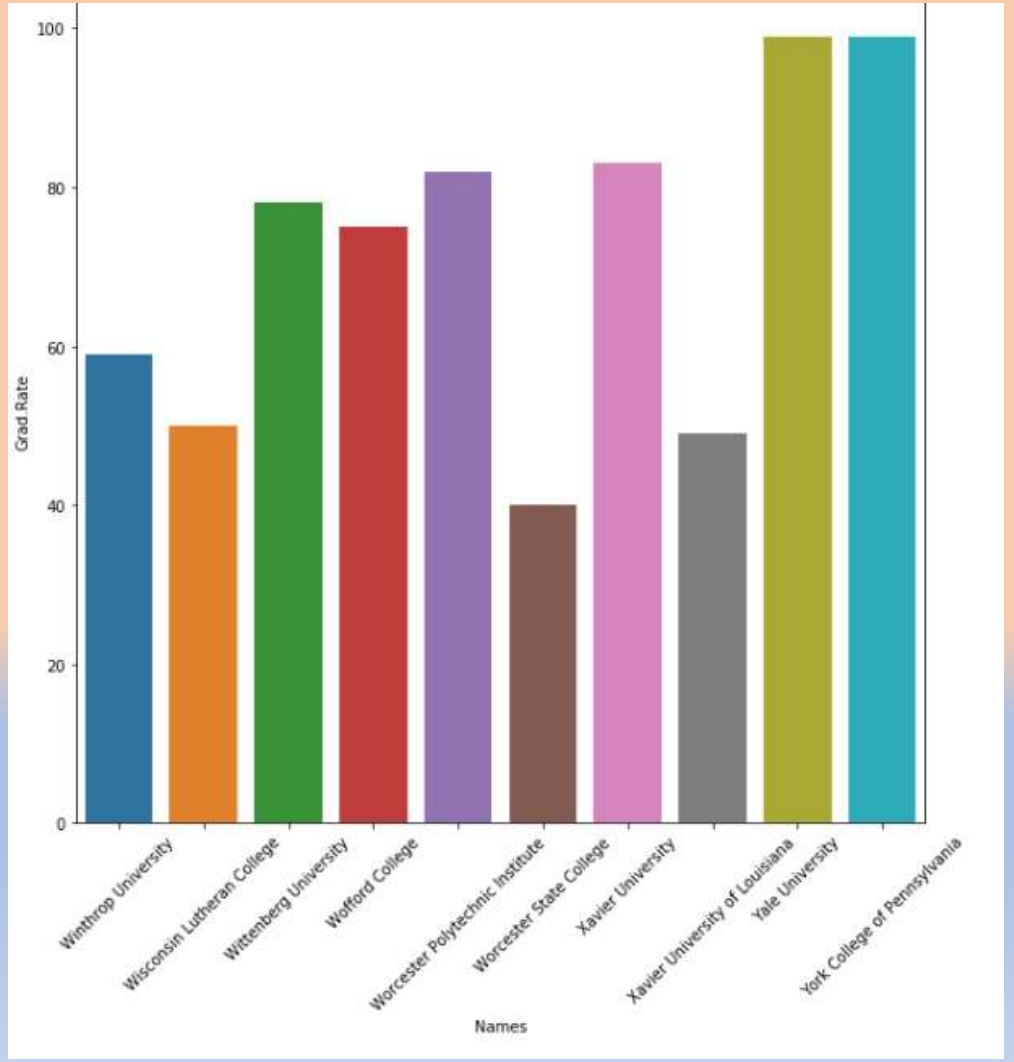
- Full time undergraduate are the most who have applied, enrolled and most of them are accepted:



- From first 10 colleges of dataset we could see Alaska Pacific University and Albright College has the lowest and highest Graduation Rate respectively:



- From last 10 colleges of dataset we could see Worcester state College and Yale University , York college of Pennsylvania has the lowest and highest Graduation Rate respectively:



Inferences on using the type of Scaling Function

- Case study mainly focuses on Principal component analysis . PCA requires/demands to standardise the dataset i.e. we need to bring the data points to the origin therefore data points needs to be subtracted from their means so that data points are centred on origin.
- To concur to the point mentioned above , StandardScaler function from sklearn library is used to scale the dataset for the case study.

Comparison between Correlation and Covariance

- Covariance of the dataset only shows whether a 2 variables are inversely or directly related i.e. it shows only the direction of relationship between 2 variables it never really the magnitude or strength of the relationship while we look at covariance matrix for the dataset there are few which '-' sign indicating inverse relationship but we can never tell about the strength of the relationship.
- While correlation matrix on the other hand shows both strength and direction of relationship between 2 variables. Correlation values of 2 variables are bounded between [-1,1] ,so values near to 1 are strongly related and directly proportional while values near to -1 are strongly related but inversely proportional. On the other values around zero indicates no significant relationship.

Outliers Before and After Scaling

- When we process the dataset we could see almost all the variables has outliers except for the variable Top25perc.
- After necessary outlier treatment and scaling we could see only 3 variables namely Top10perc,S.F.Ratio and perc.alumni are having outliers
- Since there are only few variables affected by outliers and only a sum of about only 3-4 outliers are present after scaling we could conveniently neglect the outliers present and can continue with analysis.

Covariance ,Eigen Vectors and Eigen Values

- We could see Eigen Values are constructed from Eigen Vectors and Covariance matrix. We could see first eigen value captures 5.66 units of variance from the original dataset while second eigenvalue captures 4.89 units of variance from 17 features in the dataset

```
 Eigen Values
%s [5.6625219  4.89470815 1.12636744 1.00397659 0.87218426 0.7657541
 0.58491404 0.5445048  0.42352336 0.38101777 0.24701456 0.02239369
 0.03789395 0.14726392 0.13434483 0.09883384 0.07469003]
```

Explicit form of First PC in terms of Eigen Vectors

- In generic first PC can be represented using linear combination of features and its coefficients/weights
  PC1=w11 *X1+w12* X2+w13 *X3+w14* X4.....
  where X1,X2,X3,X4... are original variables/features.
- In this scenario PC1 can be represented as linear combination of below components
  [w11 w12 w13 w14....w117]=

```
%s [[-2.62171542e-01  3.14136258e-01  8.10177245e-02 -9.87761685e-02
  -2.19898081e-01  2.18800617e-03 -2.83715076e-02 -8.99498102e-02
   1.30566998e-01 -1.56464458e-01 -8.62132843e-02  1.82169814e-01
  -5.99137640e-01  8.99775288e-02  8.88697944e-02  5.49428396e-01
   5.41453698e-03]
```

[X1 X2 X3.....X17]=['Apps' 'Accept' 'Enroll' 'Top10Perc' 'Top25Perc' 'F.Undergrad' 'P.Undergrad' 'Outstate’ 'Room.Board' 'Books' 'Personal' 'PhD' 'Terminal' 'S.F.Ratio' 'perc.alumni' 'Expend’ 'Grad.rate']

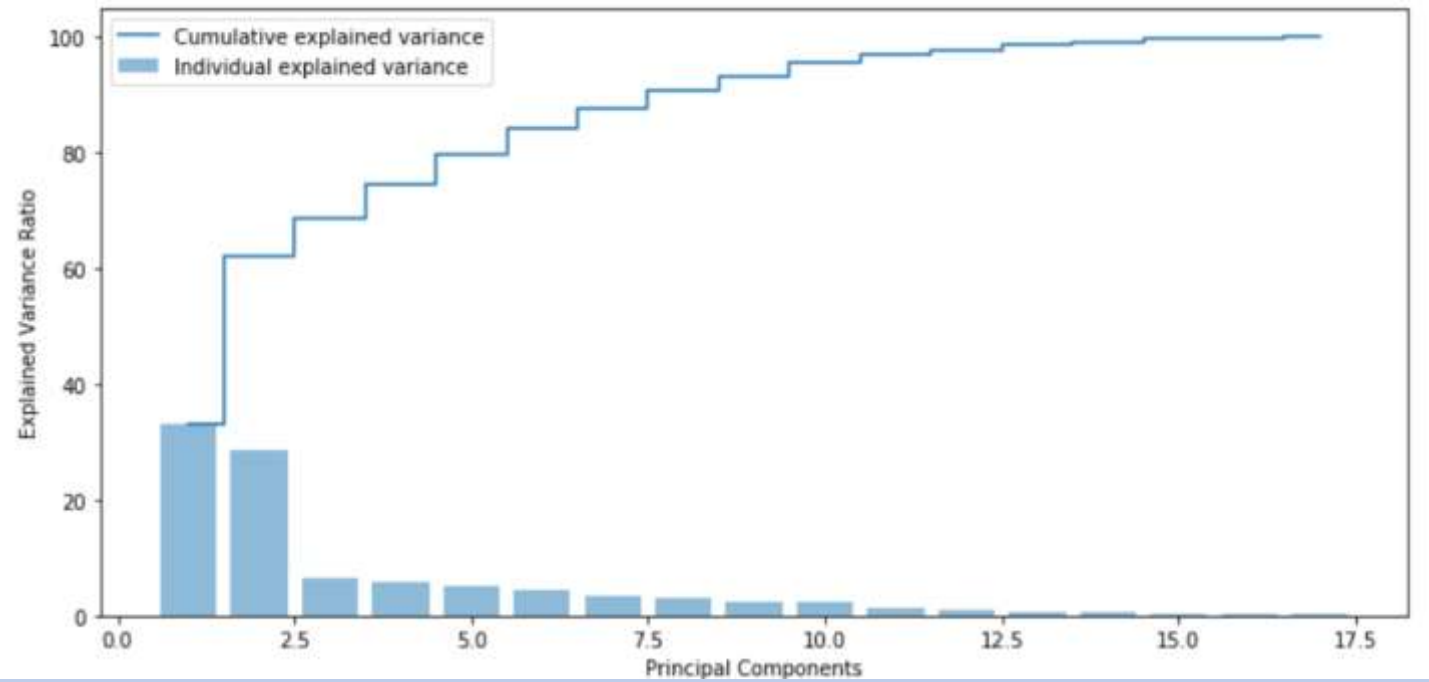
Cumulative Eigen Values and Significance

- Cumulative values of eigen values shows the percentage of variance captured by the Principal components.
- We could see first 2 PCs captures sufficiently more that 50% of the variance of the datasets, we could see almost till 6th component there is a significant amount variance capture after that variances captured are negligible

```
Cumulative Variance Explained [ 33.26608367  62.02142867  68.63859223  74.53673619  79.66062886
  84.15926753  87.59551019  90.79435736  93.28246491  95.52086136
  96.97201814  97.83716159  98.62640821  99.20703552  99.64582321
  99.86844192 100.        ]
```

Optimum Number of PCs and Eigen Vector Significance

- We could see from Scree plot after 6th component variances captured are negligible. Therefore we could only take 6 components as PCs to explain the variance of the whole dataset.



- Eigen vectors represent the directions of the new features/PCS lies in coordinate space

Business implications/Conclusion

- For the case study there are variety of parameters used for a college which could lead to indecision on which college performs best.
- There are lots of dependencies between 2 variables and there are many noise/outlier data as pointed out in boxplots and pairplots.
- PCA for this case study captures all these variations by standardizing the variables in the dataset.
- PCA improves the situation by shifting all of the data points to a different direction using eigen vectors and scaling/standardizing to represent in a different magnitude(eigen values).
- Dimension reduction also achieved –with minimal number of components at the same time capturing maximum variances when one moves along a PC.

THE END