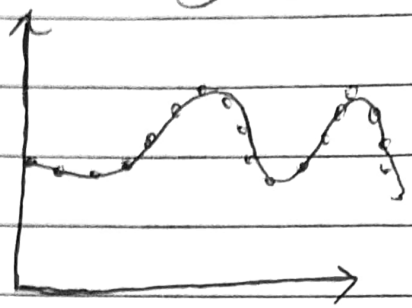
EDA pre-processing technique :

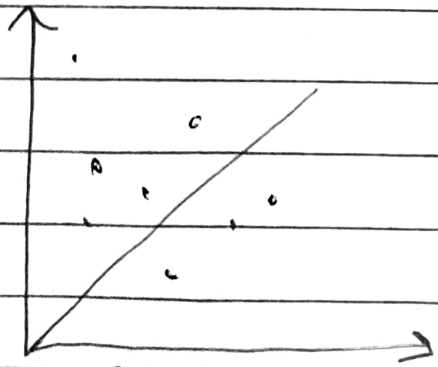
PCA (this principal component Analysis)
(A modern data Analysis)

before starting PCA concept lets understand what is overfitting and underfitting its modeling error



overfiting modeling error

it happens when excess knowledge of attributes where model tries to use all attributes (even the least important ones);



underfiting modeling error

Due to lack of enough attributes model doesnt have enough knowledg and give incoherent result

* and what is best fitting but is when model make the prediction with 0 error.

* Now we come to the PCA

to overcome the overfitting problem. we need PCA

* what it do
it reduces the overfitting problem.

why overfitting happens when there is lot of dimensions (columns) in our data set Now in order to reduce dimensions of our dataset in order to get rid of overfitting problem we need PCA (principal component Analysis).

-> Some important point.

* PCA change the orientation of dataset

* No. of pcl principal of component can be less than or equal No. of attributes

* PCA have job to compress the column and find out unnecessarily column.

- in PCA Accuracy is just go[illegible] but at best for simplicity.

DATE / /

★ whenever No. of PC is generated whom should we give priority?
=> to PC, to ascending order.

★ No. of PC should be independent to each other.
and PC have to capture [illegible]cence of column.

★ let take example in order to unders PCA.

(1) Eigen values :- it going to capture the essen of the dataset

(2) Eigen values / vector :- its going to capture direction for us.

Use of PCA

(1) identify relation between columns

(2) reduce columns

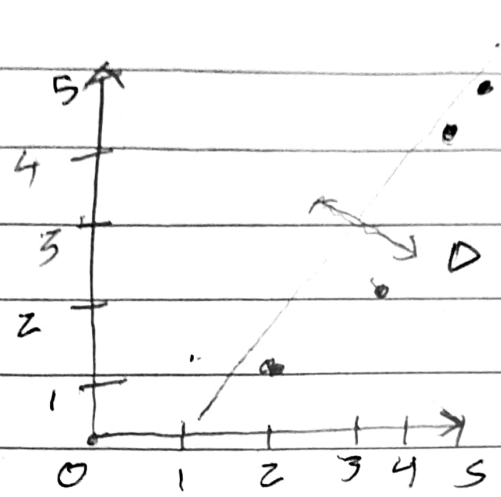
(3) visualize data in 2D

PAGE No.
DATE / /

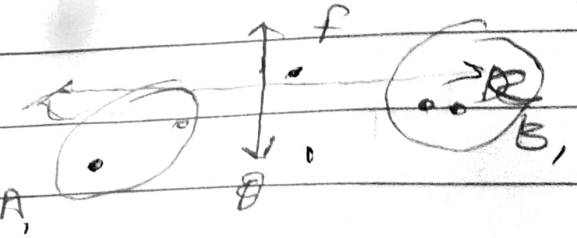
* data in PCA should be in Linear fashion among the column then only PCA will perform well otherwise not.

Suppose we dataset

A	B
1	2
5	6
2	3
4	5



So we have dataset in linear fashion so we can fit line to calculate variance (line c) and since we have 2nd column too we can put more line (note:- the line we are putting we assume as axis) and we can calculate more variance now lets as assume c as PC_1 and D PC_2 and if we have more columns and so on.

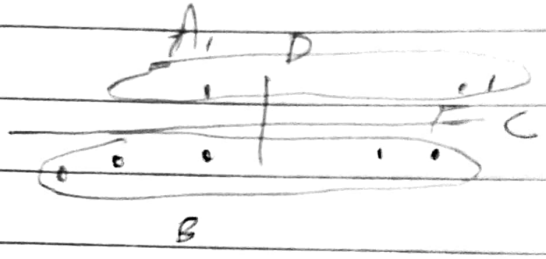


So the most data on circle is deleva
for line C and datapant with them.

PC_1

AB	Influence	in numbers
A_1	high	10
B_1	medium	5
F	low	0.2

Similar for D



PC_3

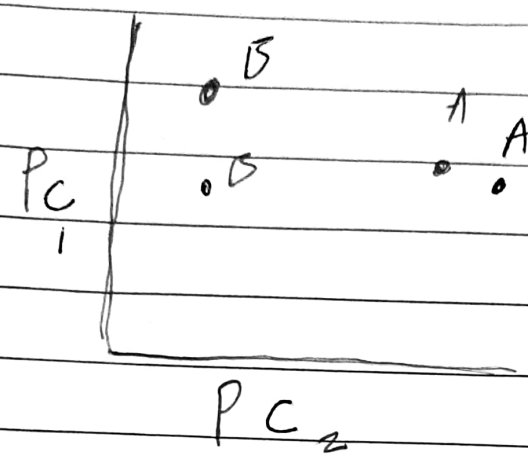
AB	influence	in number
A_1	medium	5
B_1	high	10
F	low	0.5

A PC_1 Score = $(1 * 10) + (5 * 5)$

$P_{C_1} = = 12$

Similar for PC_2

$= 6$



conclusion:- this is basic under standing for PCA we will be using mathematical formula for this calculating PCA.