There are two parts of Unsupervised Learning

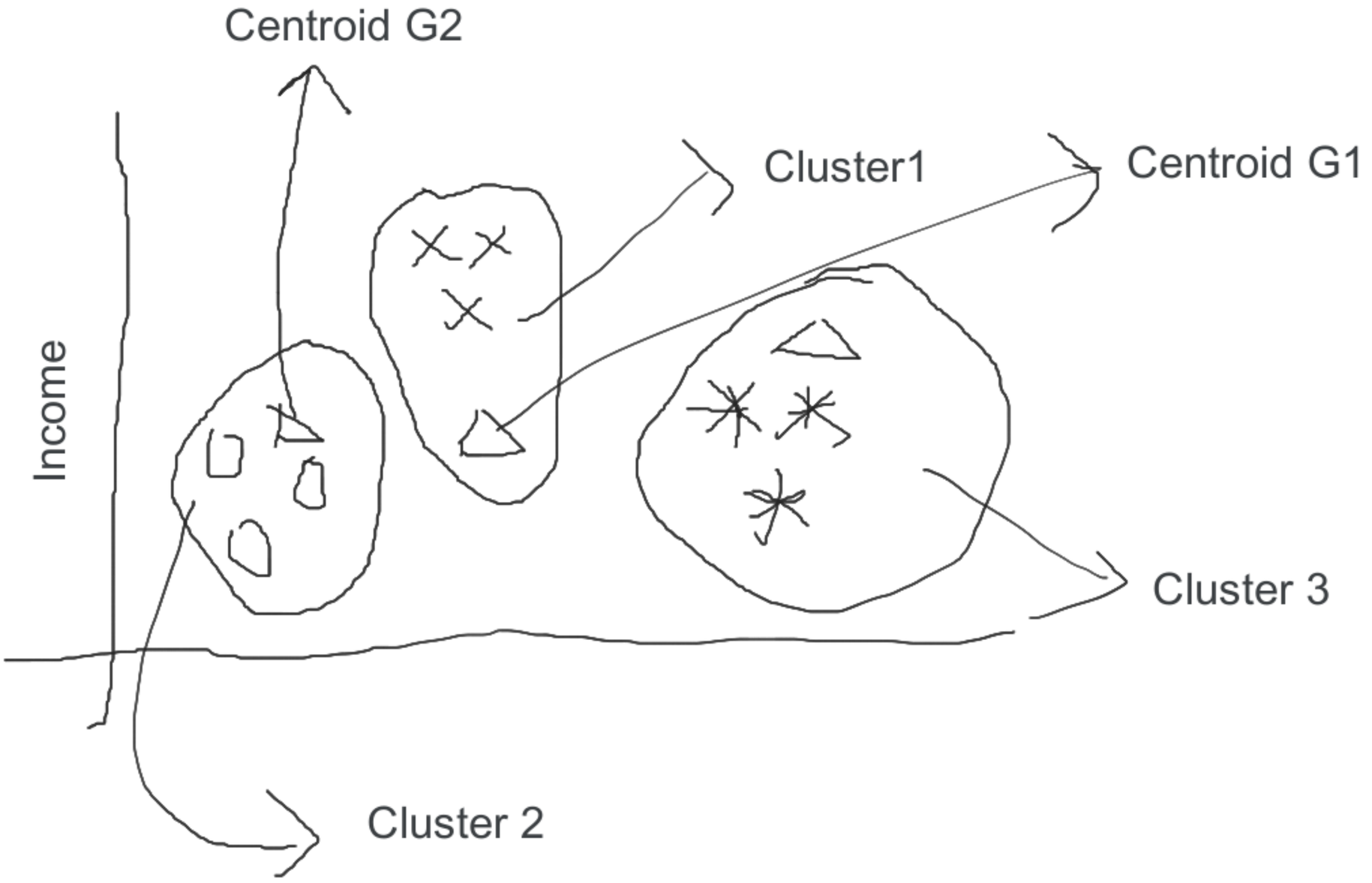
1. KMeans Clustering
2. Principal Component Analysis

## K Means Clustering

Important Case Studies

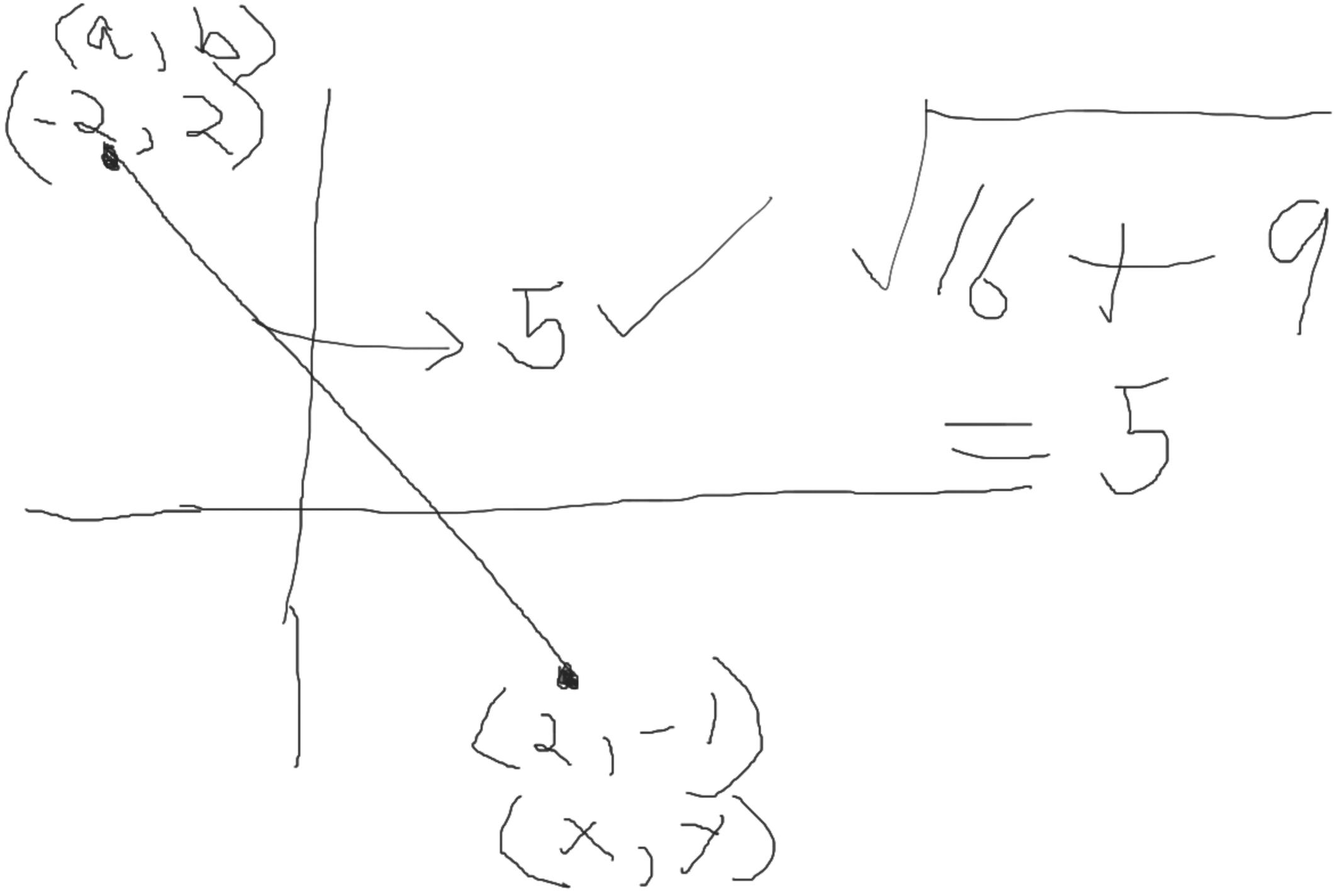
Walmart
Credit Card Segmentation

Working


Centroid G2
Cluster1
Centroid G1
Income
Cluster 3
Cluster 2

# Euclidean Distance Formula

$$d(x,y)(a,b) = \sqrt{(x-a)^2 + (y-b)^2}$$


(a, b)
(-2, 2)
5
(2, -1)
(x, y)
√16 + 9
= 5

n_clusters=3 ⟹ K

V.Imp

How to decide how many clusters? → how to choose the value of K

1. Business Objective - According to the client you will categorize/ clusterify/create groups. 100% of the time, follow the business objective.

2. Elbow Method

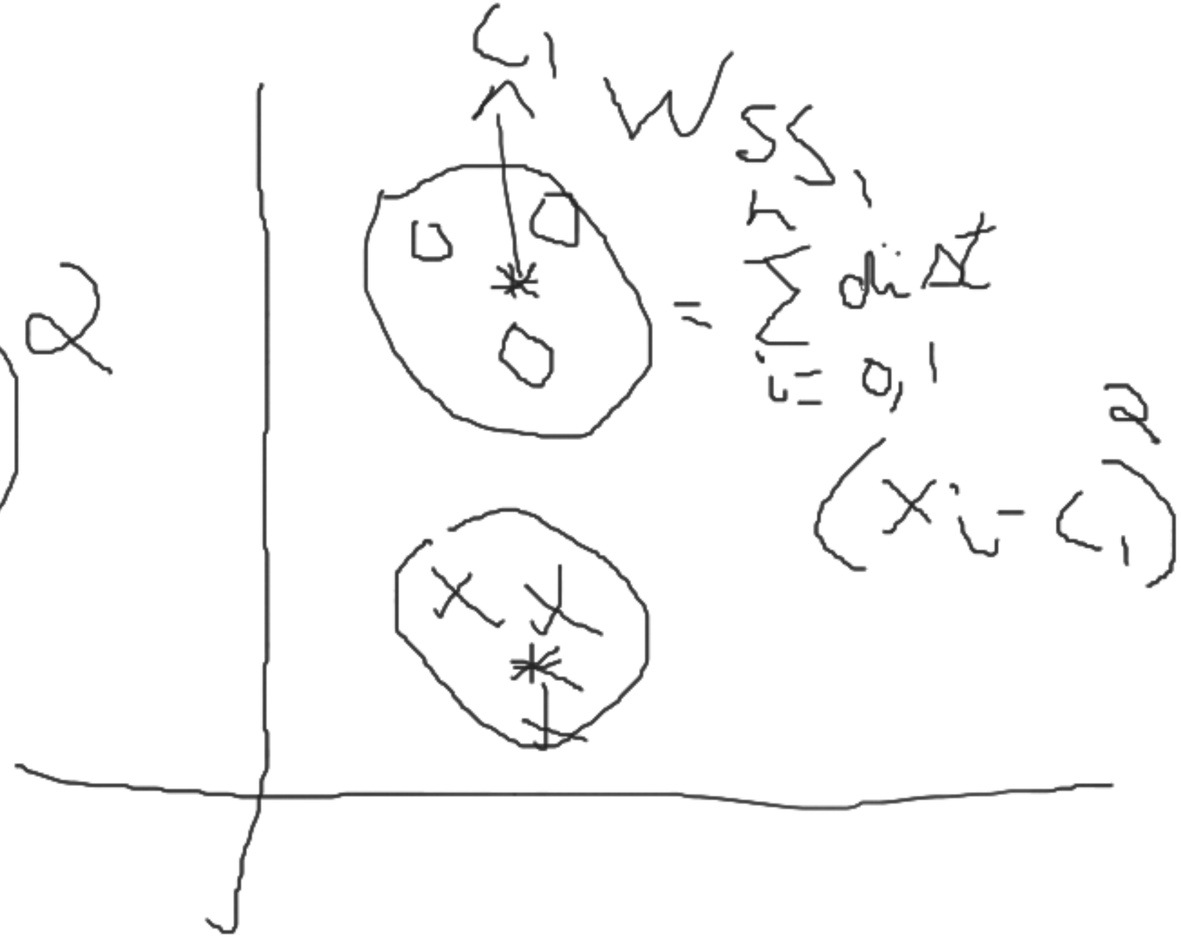
3. Sillhoutte Method

## Elbow Method ✓ V.Imp

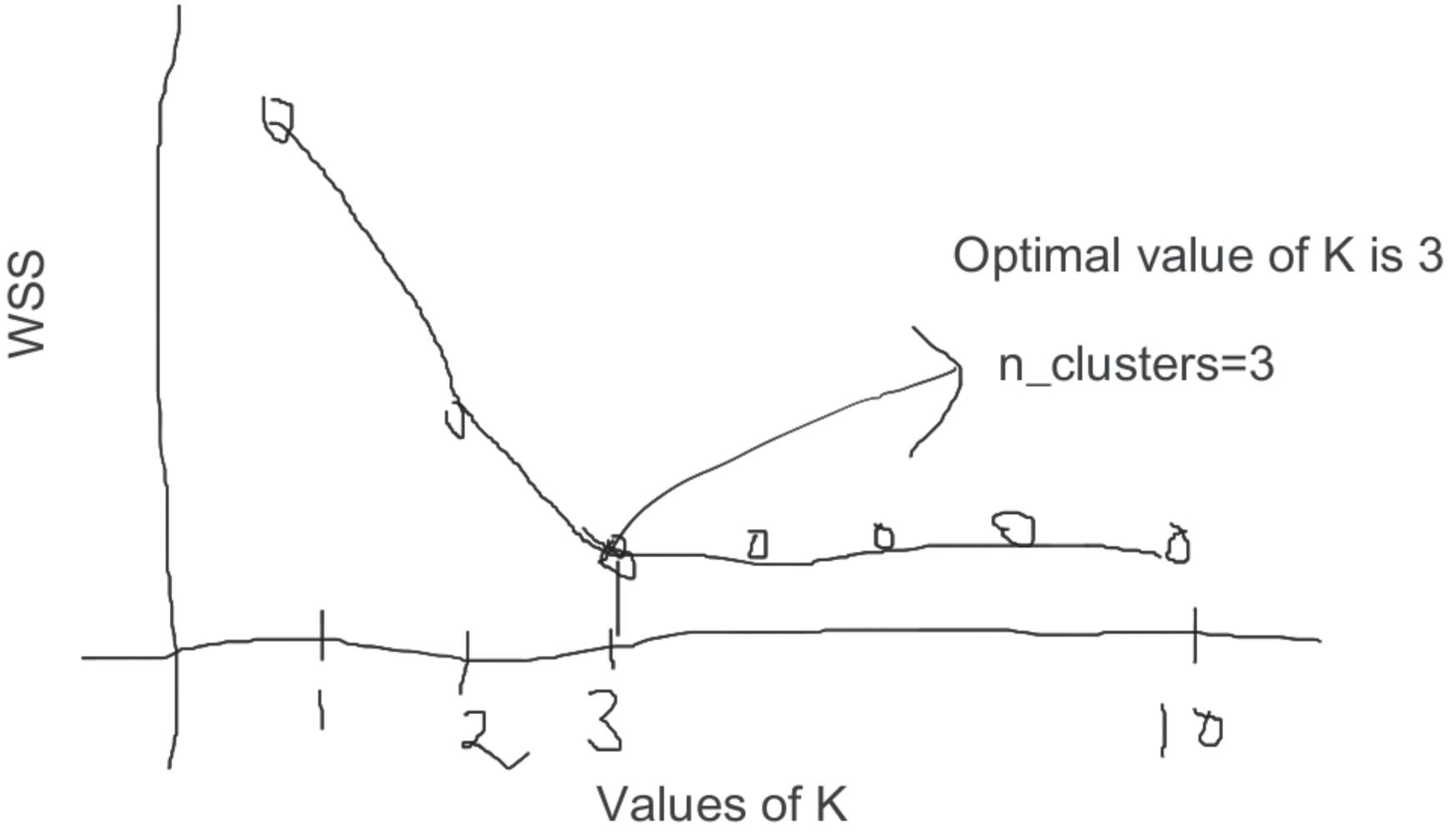
## Within Sum of Squares

WSS = WSS1 + WSS2+.....+WSSn

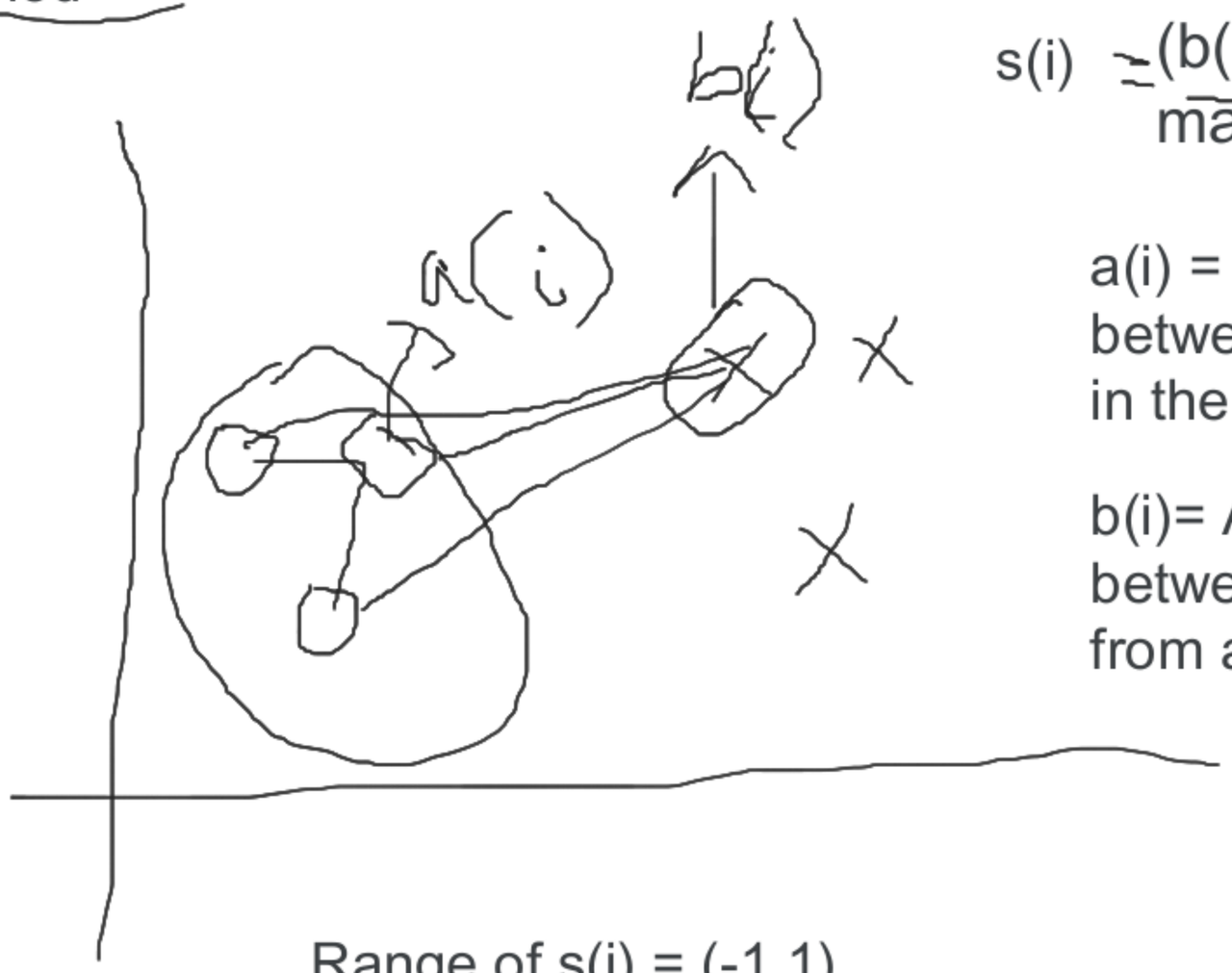
$$WSS1 = \sum_{i=0,1}^{n} dist\left(x_i - C_1\right)^2$$



Plotting of WSS values against values of K



# Sillhoutte Method




$$s(i) = \frac{(b(i)-a(i))}{max(b(i),a(i))}$$

a(i) = Average distance between every data point in the same cluster.

b(i)= Average distance between every data point from a different cluster

Range of s(i) = (-1,1)

s(i) = Sillhoutte Coefficient

If value is closer to -1 that means the data point is kind of dissimilar to the cluster or it does not belong to that cluster.

If value is closer to +1 that means the point belongs to the cluster.

# Scaling of Data

MinMaxScaler → Range from 0 to +1

StandardScaler →

Standard Scaler calculates Z Scores

$$z = \frac{X - \bar{X}}{SD}$$

X → set of discrete values of x

$\bar{X}$ → Mean of X

$$SD = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

# Need Of Scaling

| Income(USD) | Cluster | Income(Rs.) |
|---|---|---|
| 6000 | 0 | 420000 |
| 7000 | 0 | 490000 |
| 4000 | 2 | 280000 |
| 5000 | 1 | 350000 |
| 5500 | 1 | 375000 |