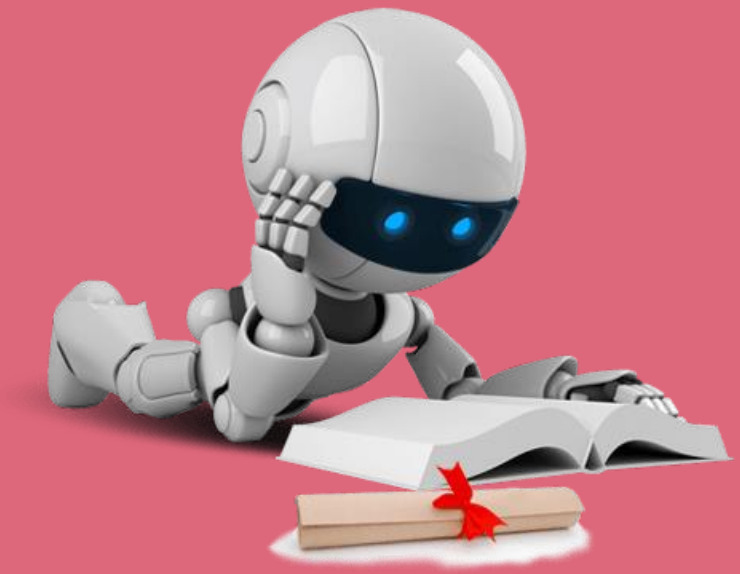
دوره‌ی آموزشی «علم داده»

Data Science Course



جلسه بیست و سوم - (بخش اول)

نکاتی در خصوص خوشه‌بندی و استانداردسازی ویژگی‌ها

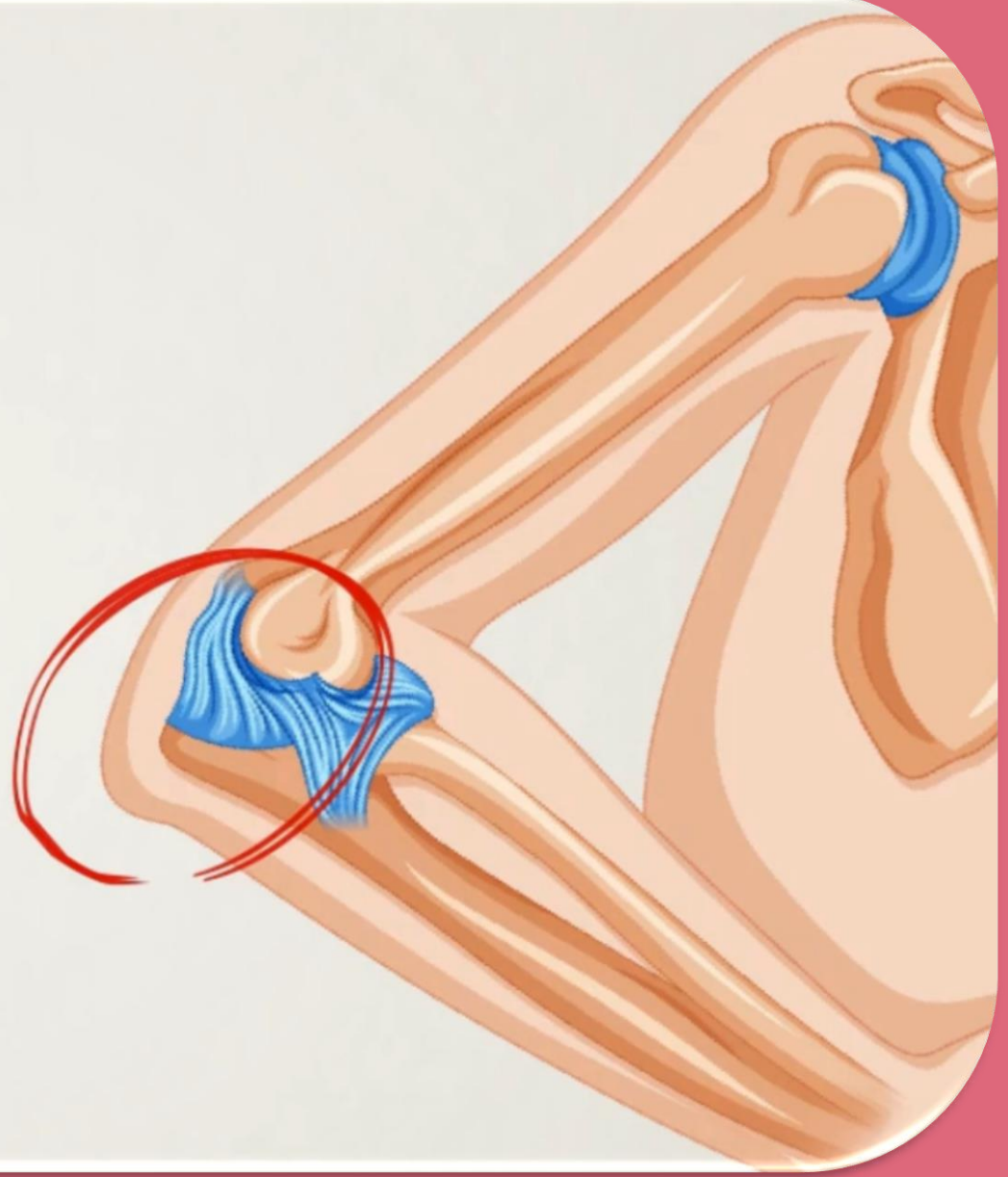
مدرس: محمد فزونی

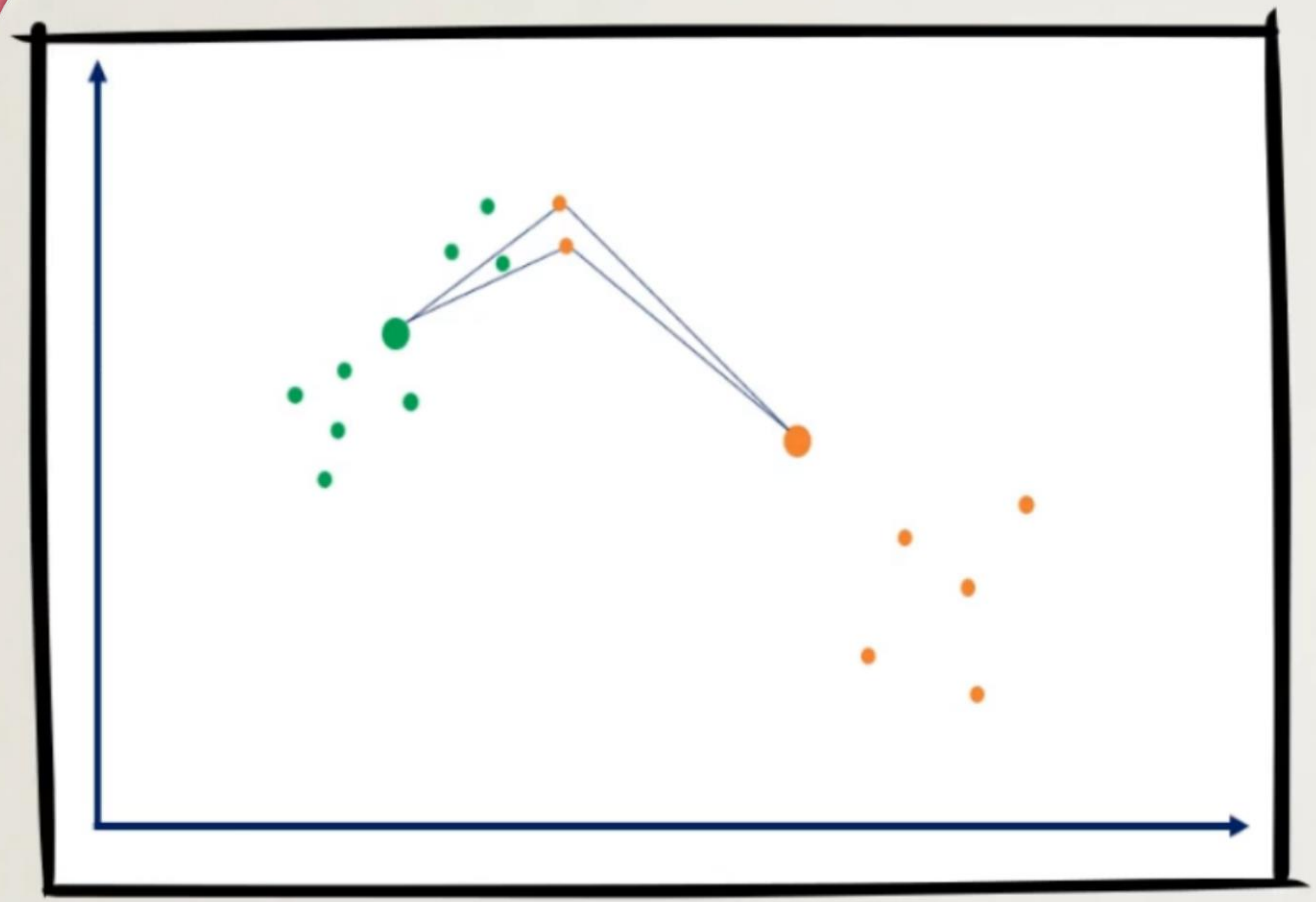
عضو هیئت علمی دانشگاه گنبدکاووس

SELECTING THE NUMBER OF CLUSTERS

CRITERION?

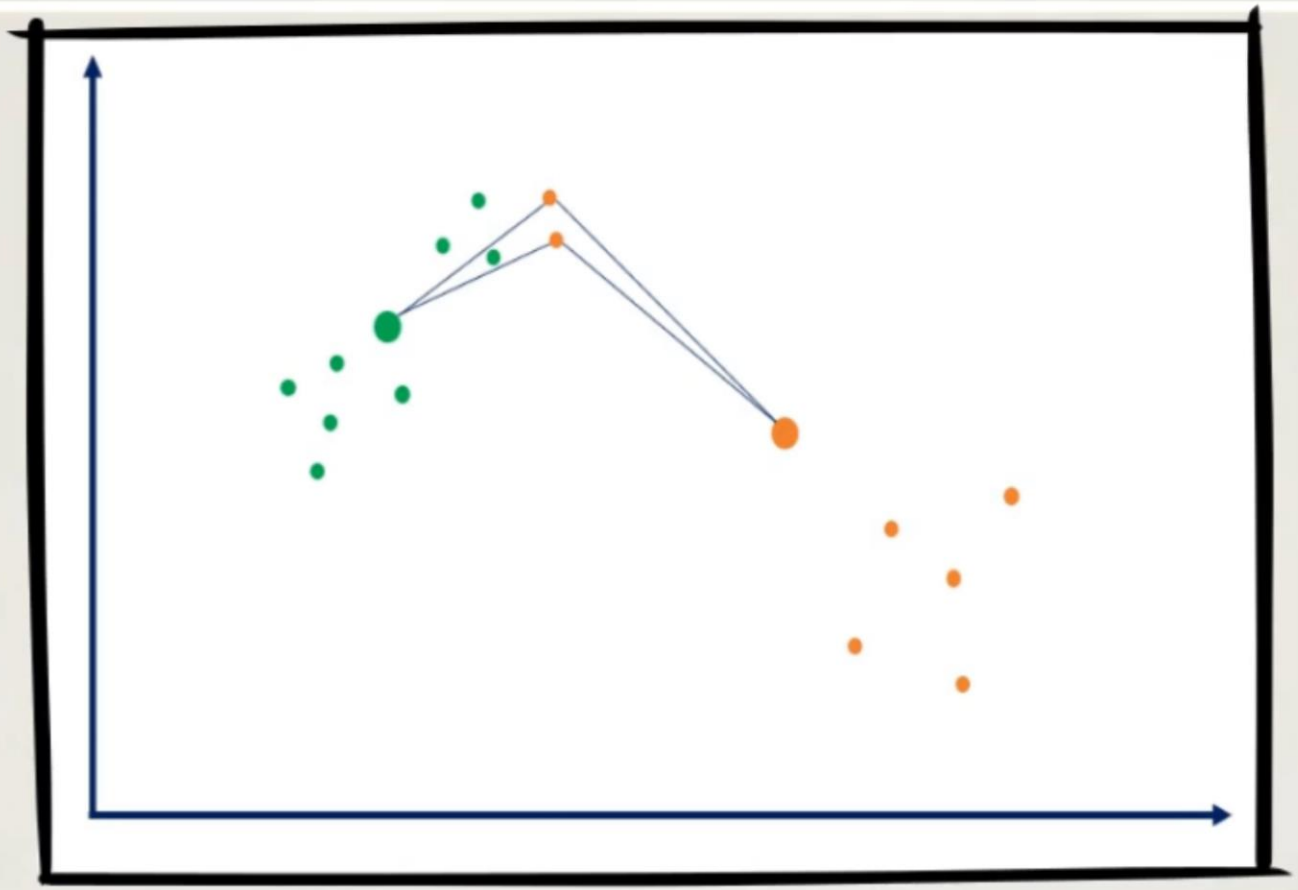
THE ELBOW METHOD





1) minimizing the distance between points in a cluster

2) maximizing the distance between clusters



1) minimizing the distance between points in a cluster

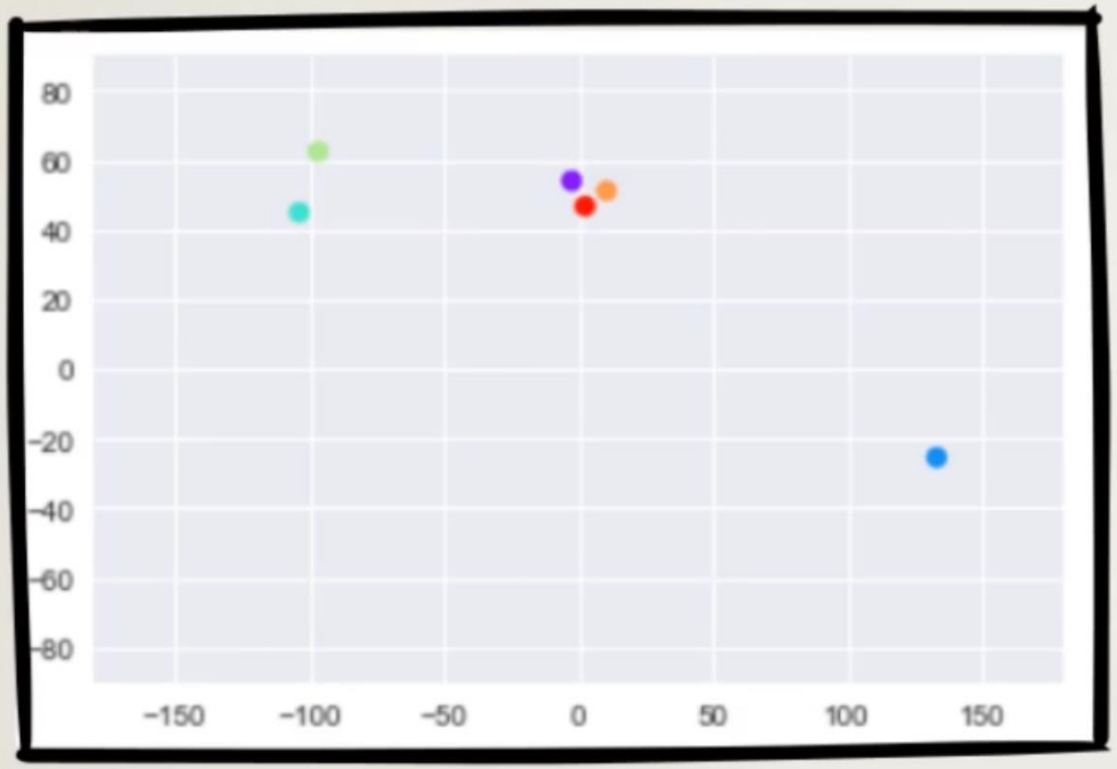
=

2) maximizing the distance between clusters

distance between points in a cluster

'within-cluster sum of squares', or WCSS

WCSS



observations: 6

clusters: 6

WCSS = 0

WCSS

observations: 1,000,000

clusters: 1,000,000

WCSS = 0 = min

observations: 1,000,000

clusters: 1

WCSS = max

MIDDLE GROUND?

WCSS

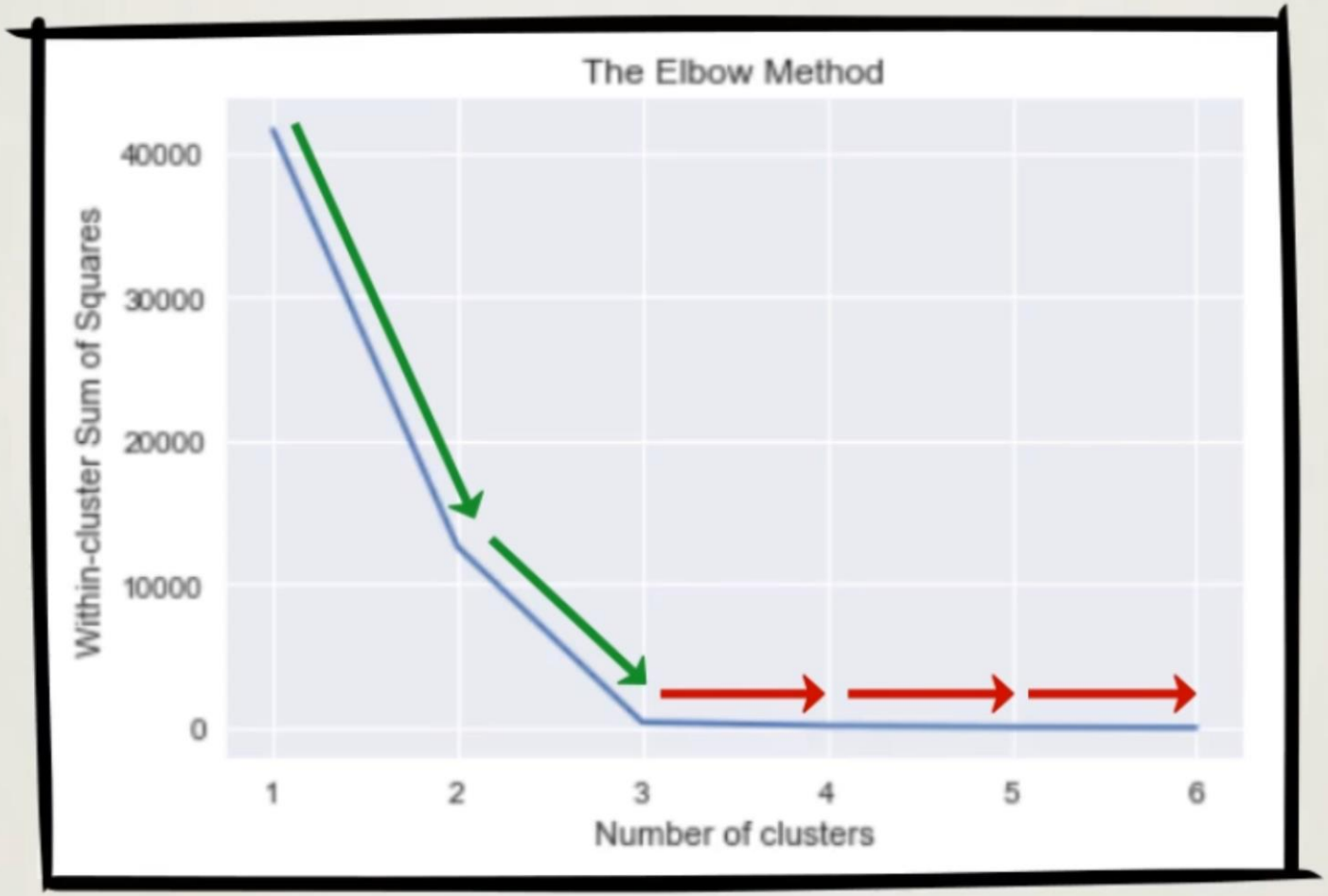
MIDDLE GROUND?

observations: N

clusters: SMALL

WCSS = LOW

THE ELBOW METHOD



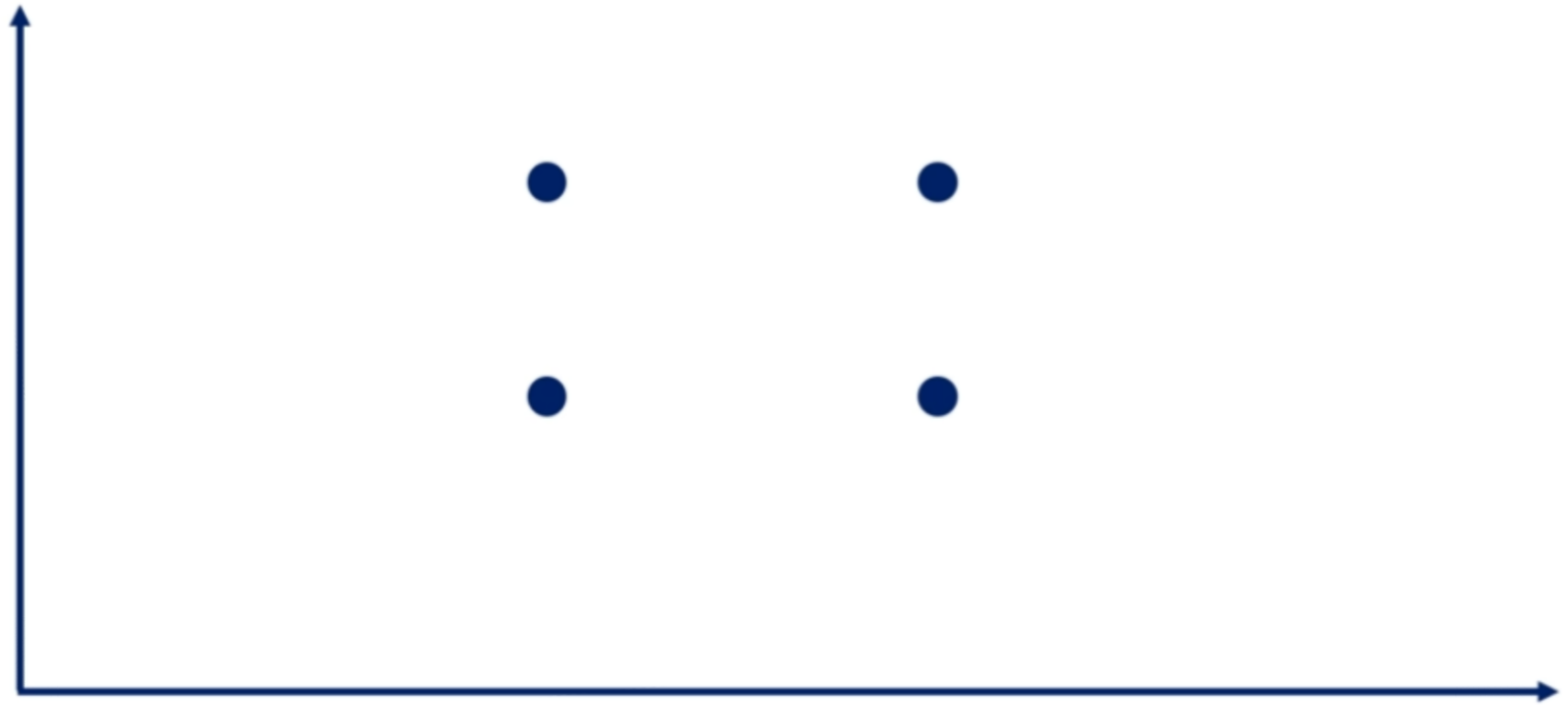
$$\mathrm{WCSS}(k) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \text{cluster } j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2,$$

where $\bar{\mathbf{x}}_j$ is the sample mean in cluster j

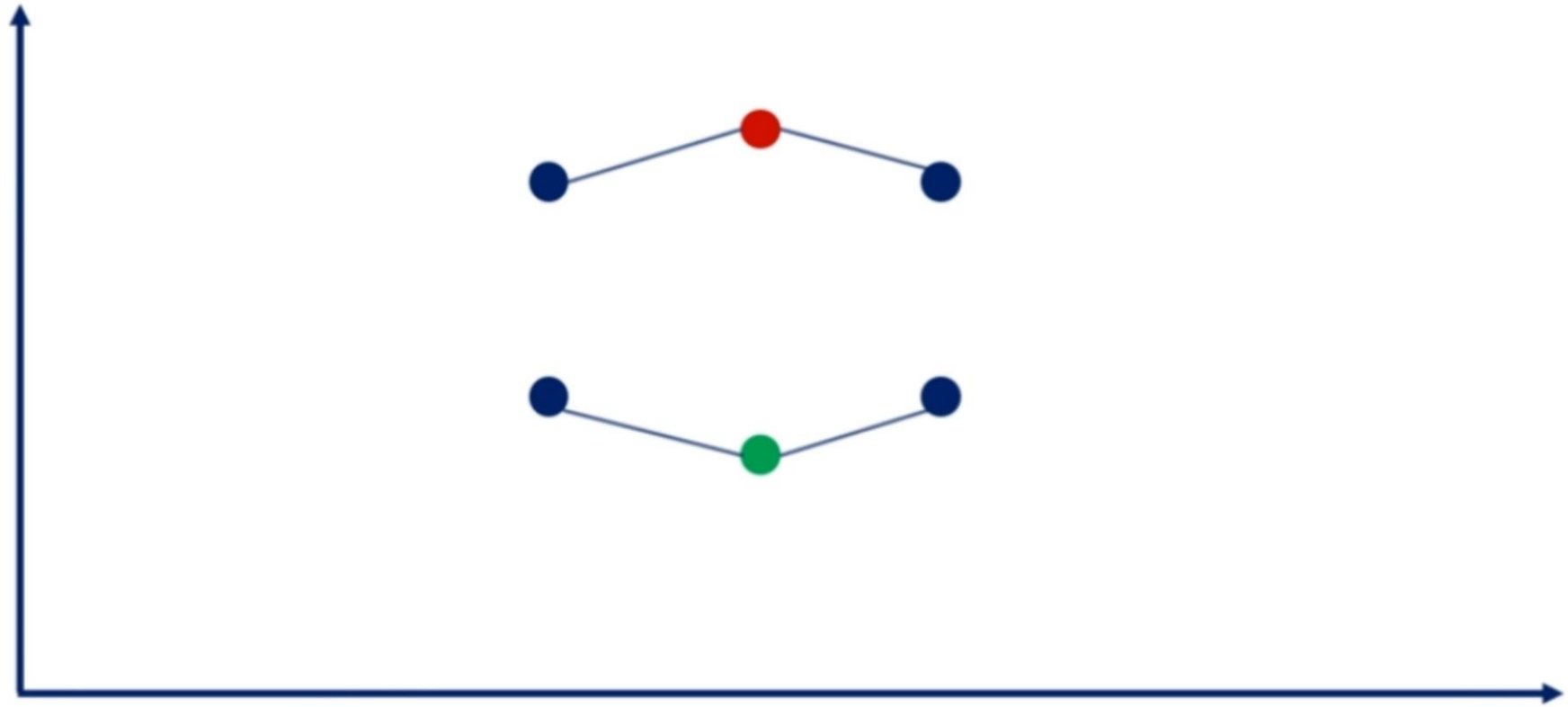
Let’s re-visit

PROS AND CONS OF K-MEANS

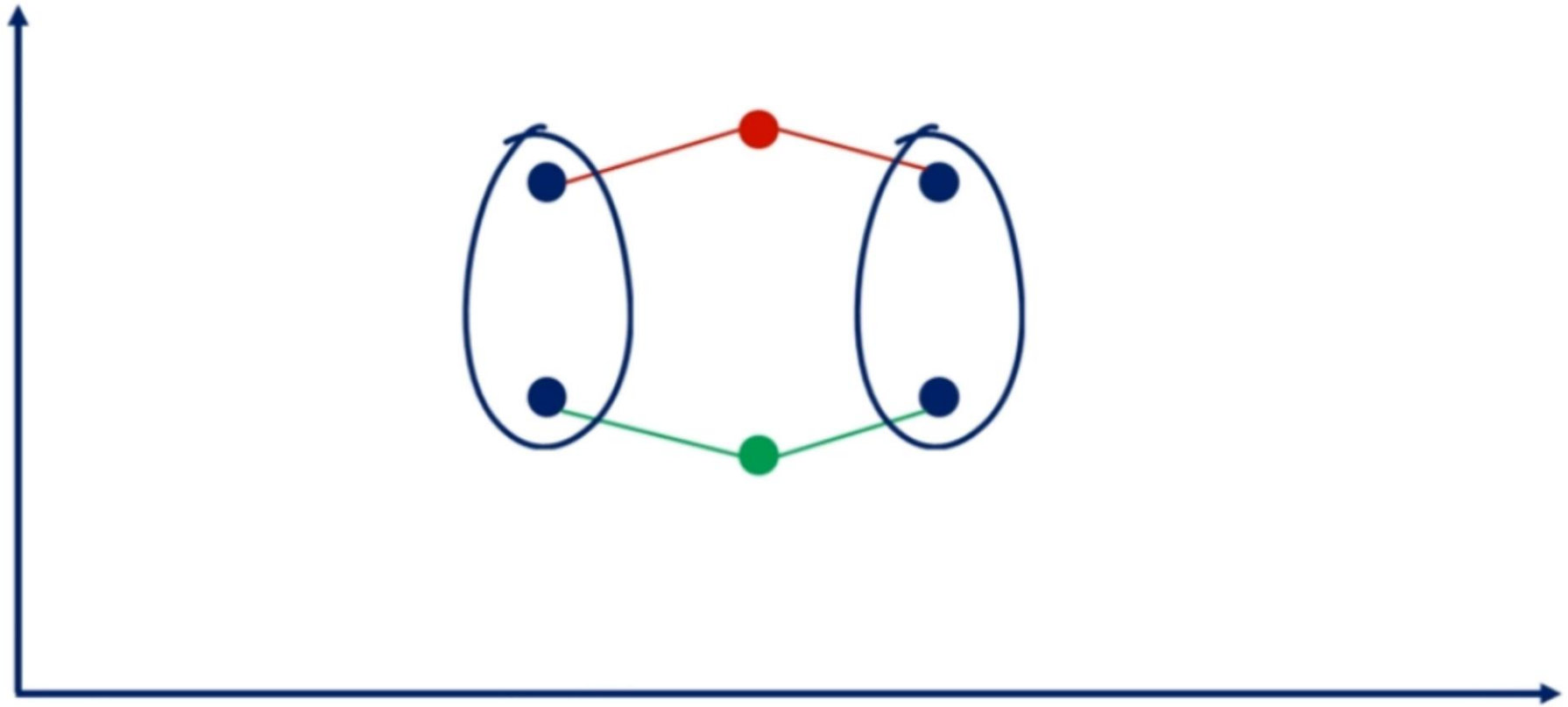
PROS AND CONS OF K-MEANS



PROS AND CONS OF K-MEANS



PROS AND CONS OF K-MEANS

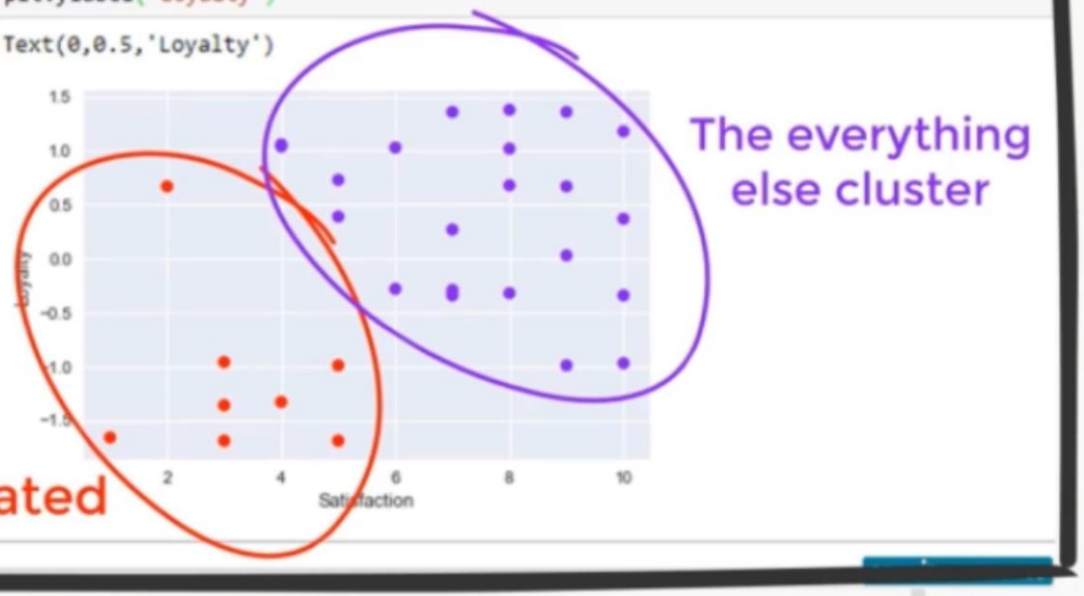


What's the remedy

This algorithm ensures a smarter initialization of the centroids and improves the quality of the clustering. Apart from initialization, the rest of the algorithm is the same as the standard **K-means** algorithm. That is **K-means++** is the standard **K-means** algorithm coupled with a smarter initialization of the centroids.

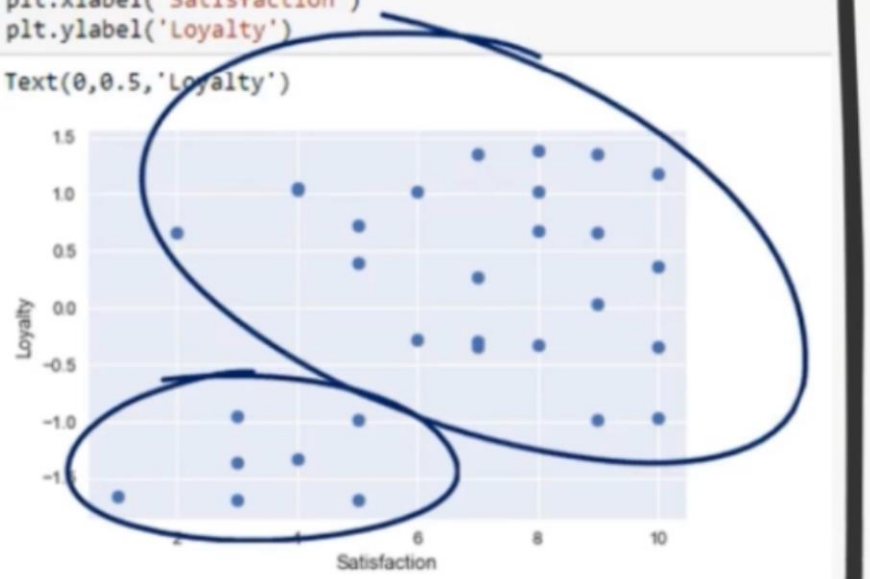
Spherical

```
In [14]: plt.scatter(clusters_new['Satisfaction'],clusters_new['Loyalty'],c=clusters_new[
         plt.xlabel('Satisfaction')
         plt.ylabel('Loyalty')

Out[14]: Text(0,0.5,'Loyalty')
```



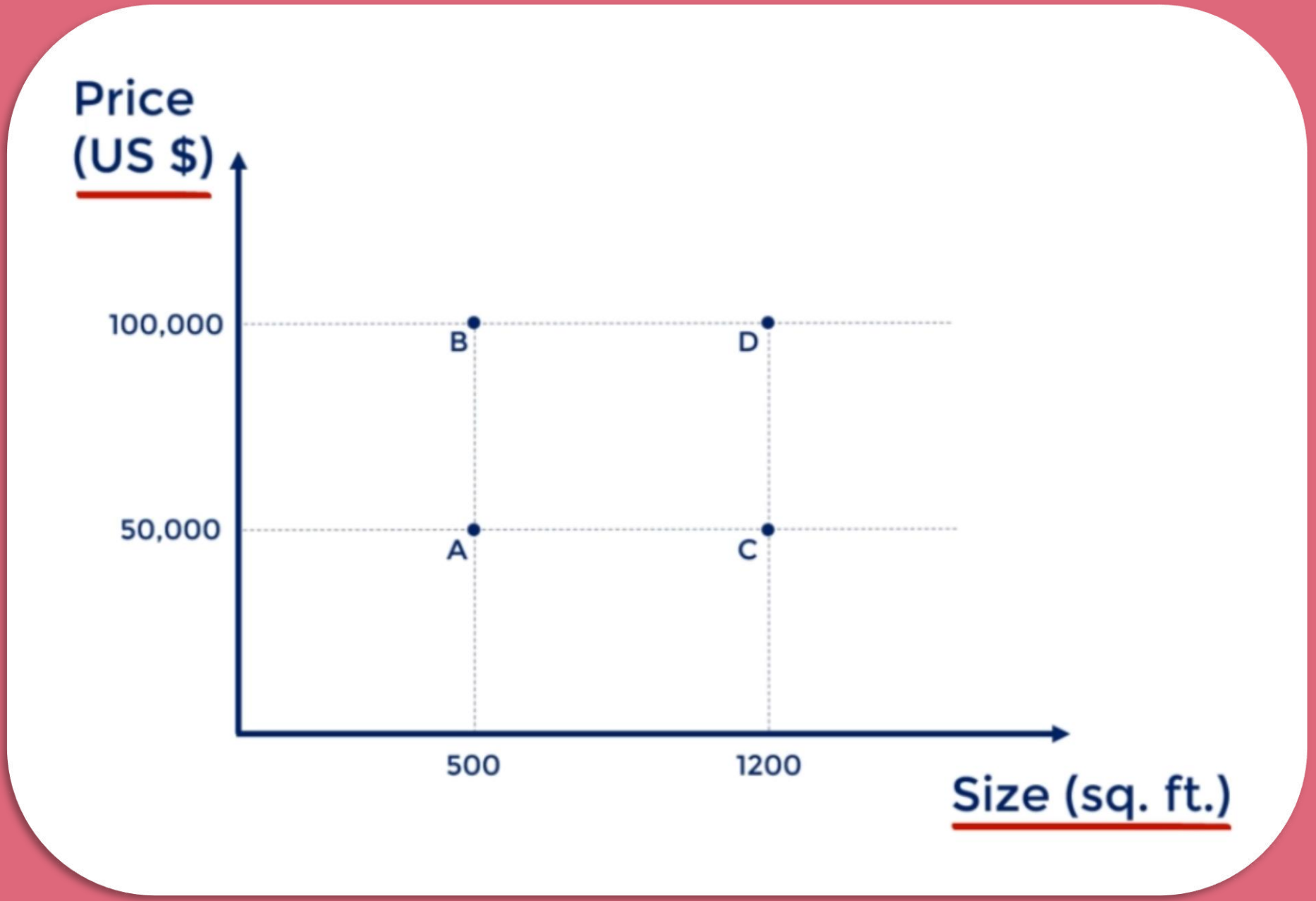
Elliptic

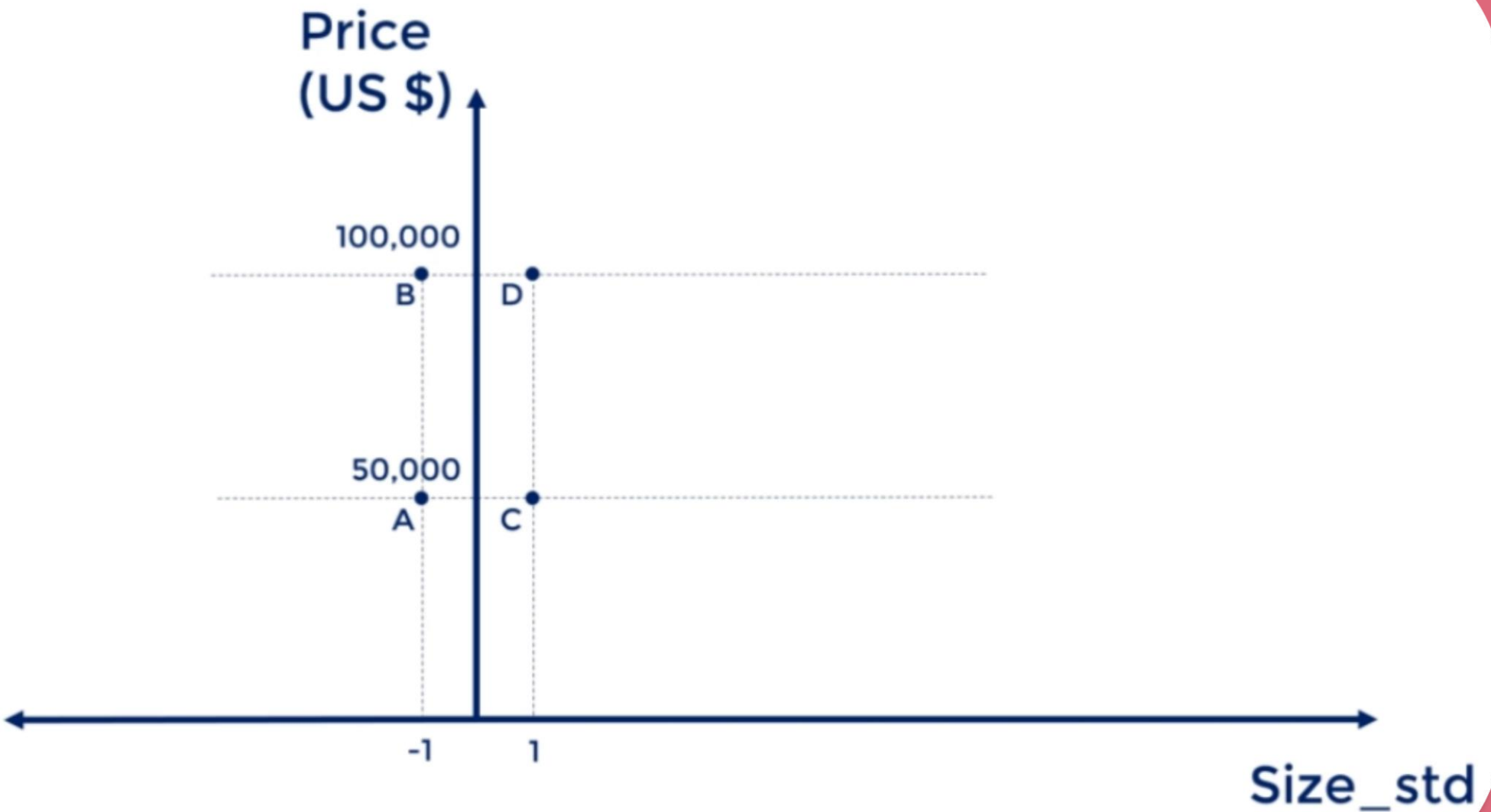
```
plt.scatter(data['Satisfaction'],data['Loyalty'])
plt.xlabel('Satisfaction')
plt.ylabel('Loyalty')

Text(0,0.5,'Loyalty')
```

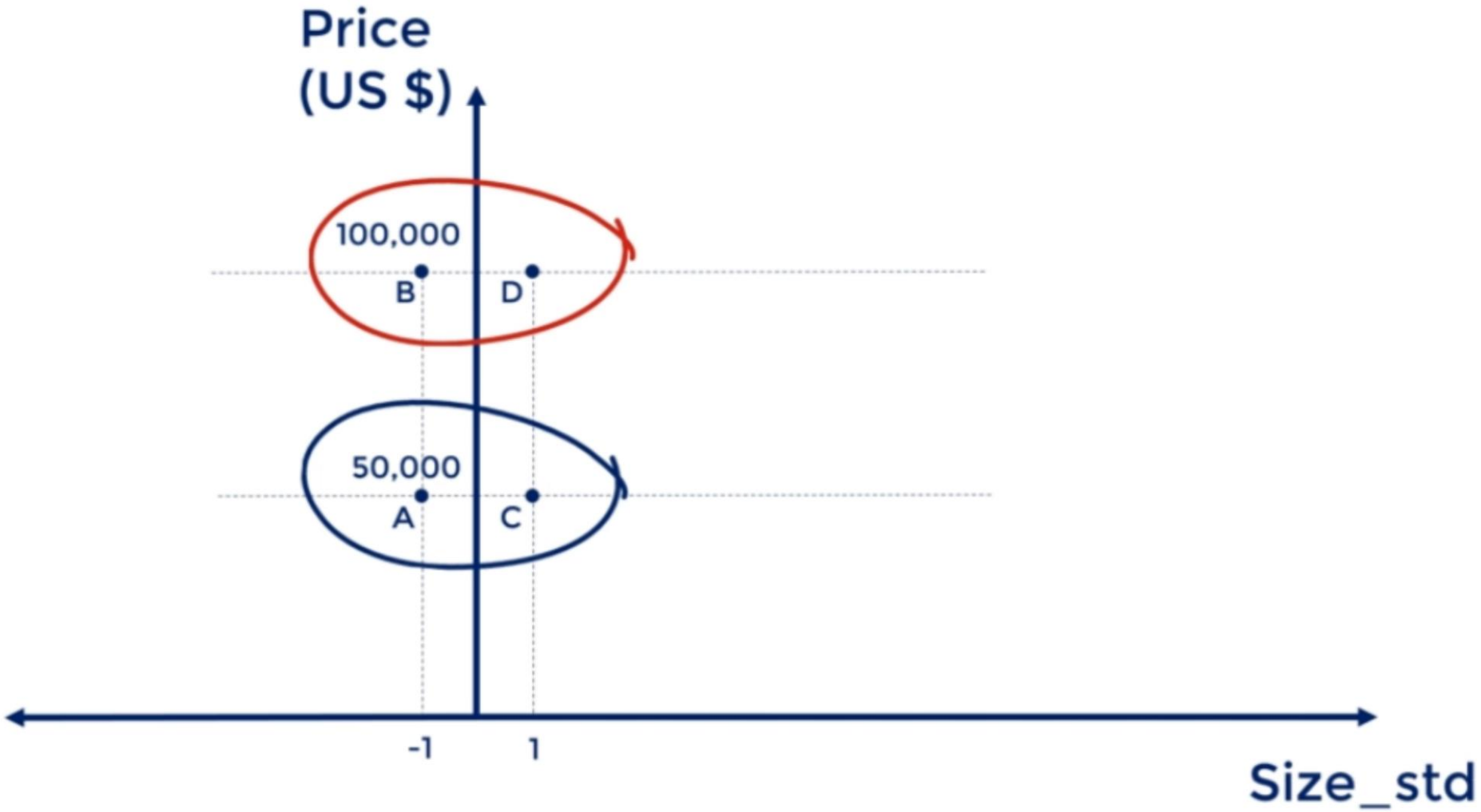


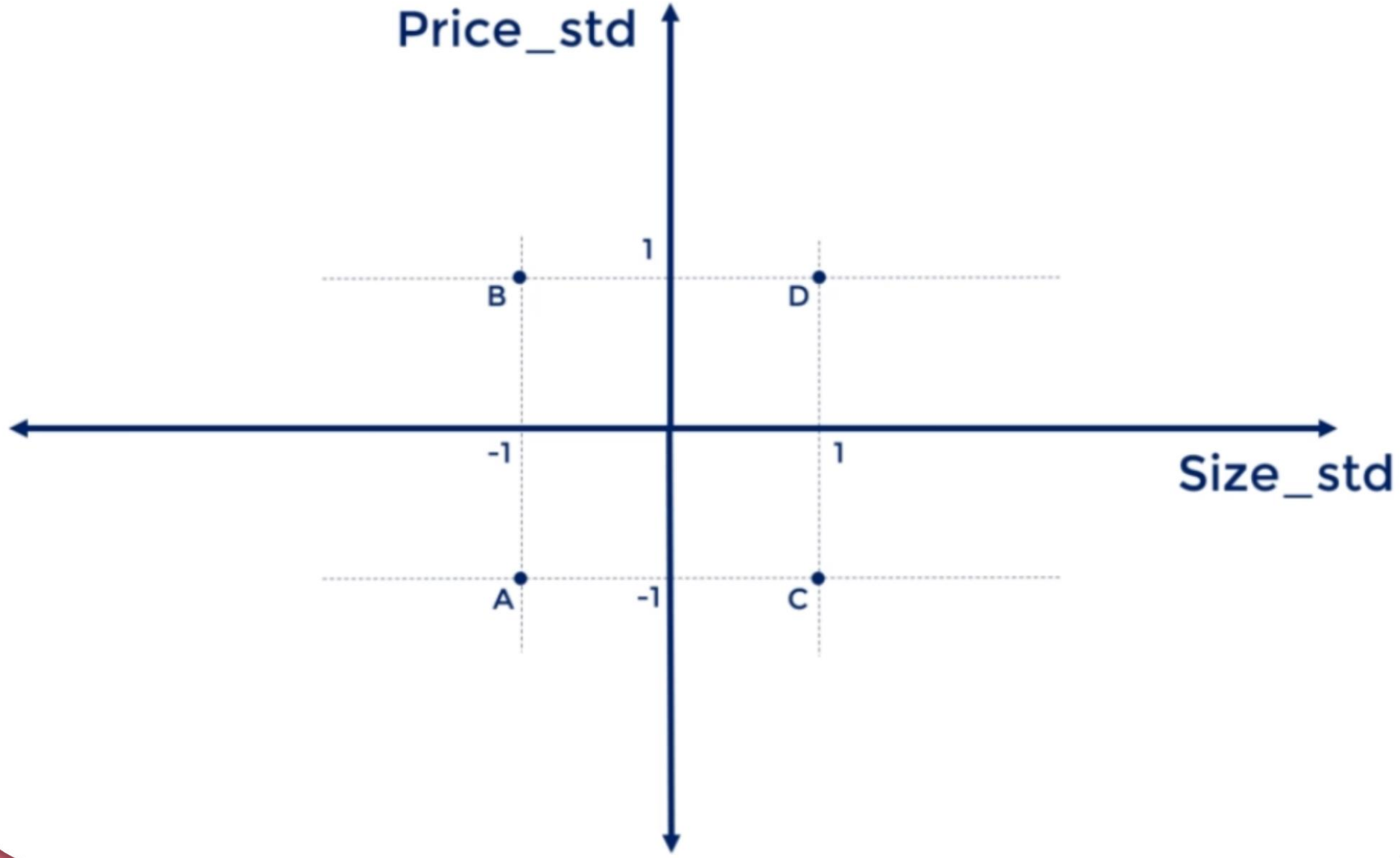
CONS	REMEDIES
1. We need to pick K	1. The Elbow method
2. Sensitive to initialization	2. k-means++
3. Sensitive to outliers	3. Remove outliers
4. Produces spherical solutions	
5. Standardization	

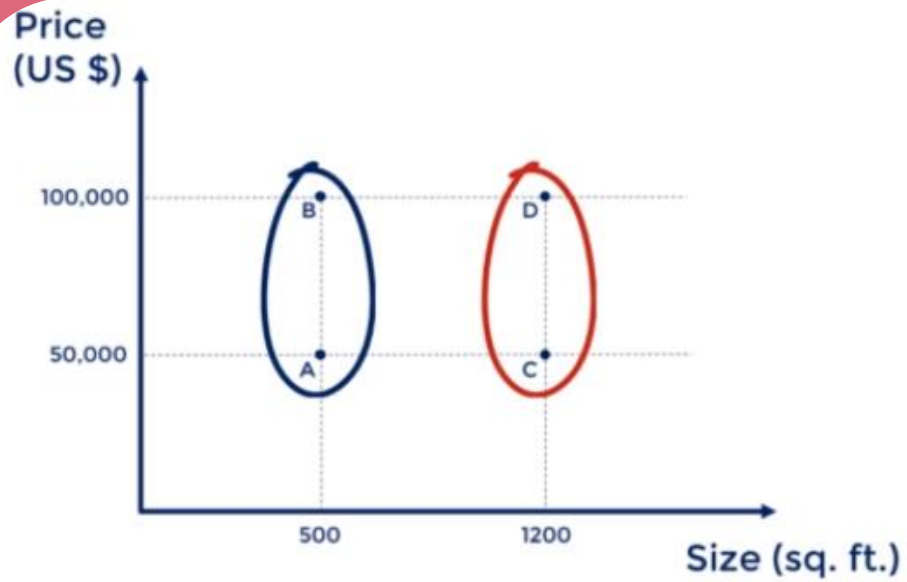
TO STANDARDIZE OR TO NOT STANDARDIZE?

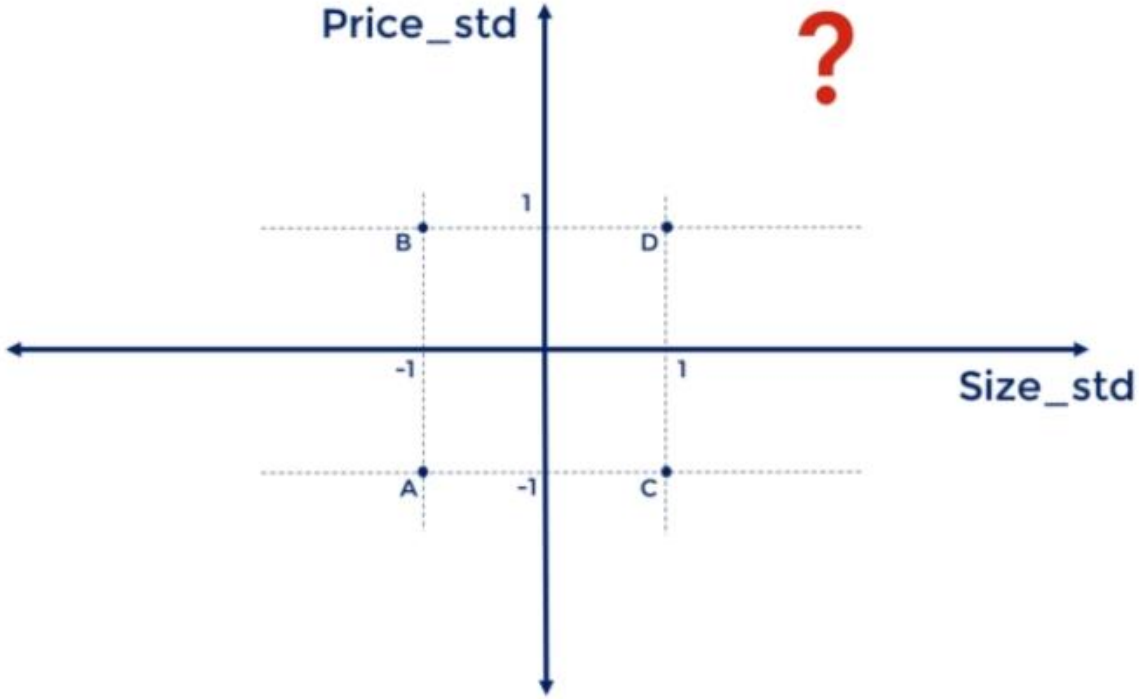

Price
(US $)
100,000
50,000
B
D
A
C
500
1200
Size (sq. ft.)

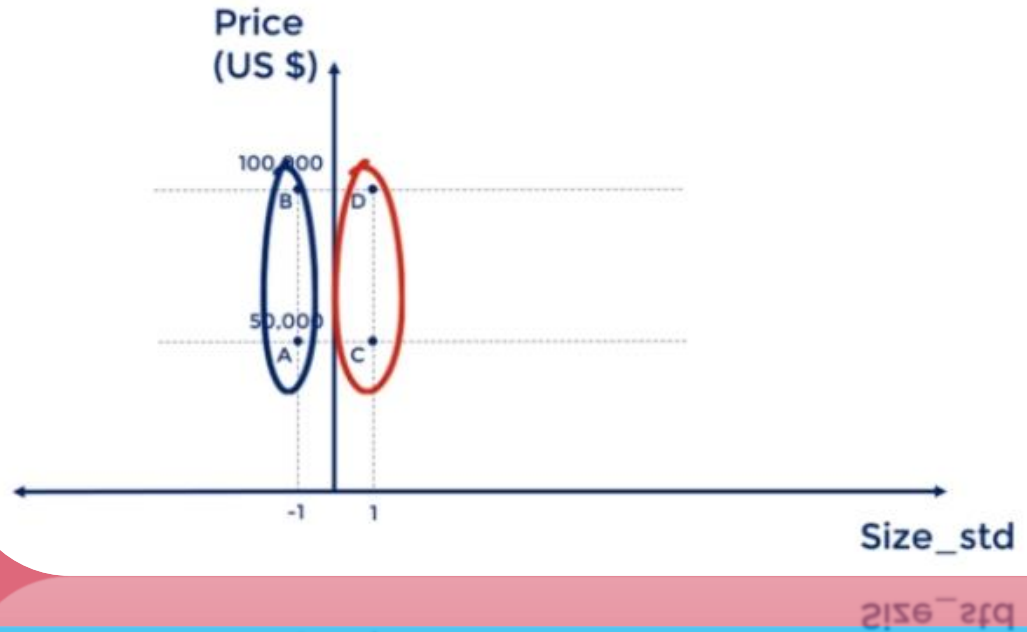

Price
(US $)
100,000
50,000
B
D
A
C
-1
1
Size_std

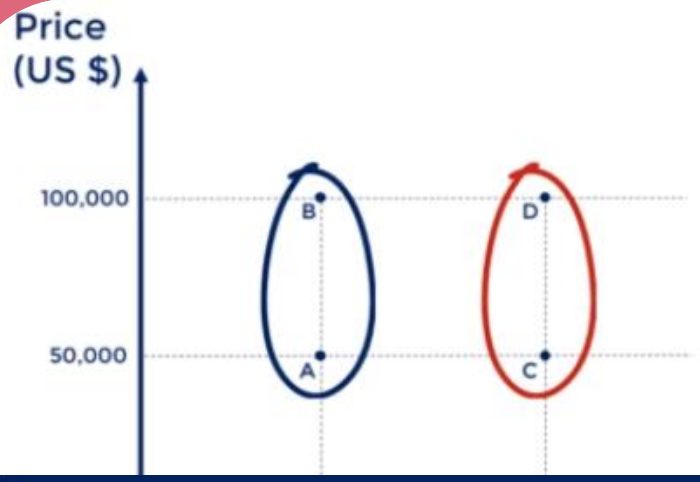

Price
(US $)
100,000
B
D
50,000
A
C
-1
1
Size_std

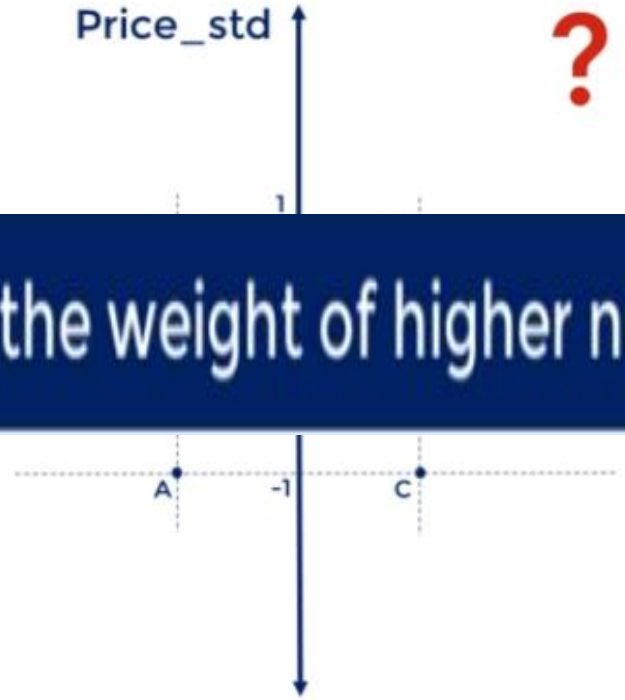

Price_std
Size_std
1
-1
1
-1
B
D
A
C


Price
(US $)
100,000
50,000
B
A
D
C
500
1200
Size (sq. ft.)

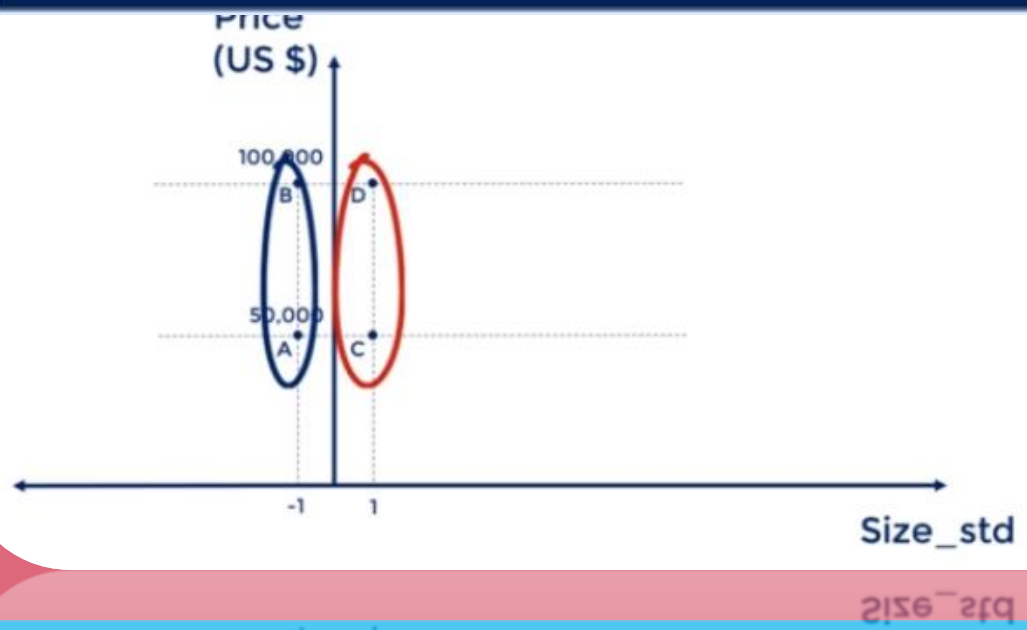

Price_std
?
1
B
D
-1
1
A
-1
C
Size_std

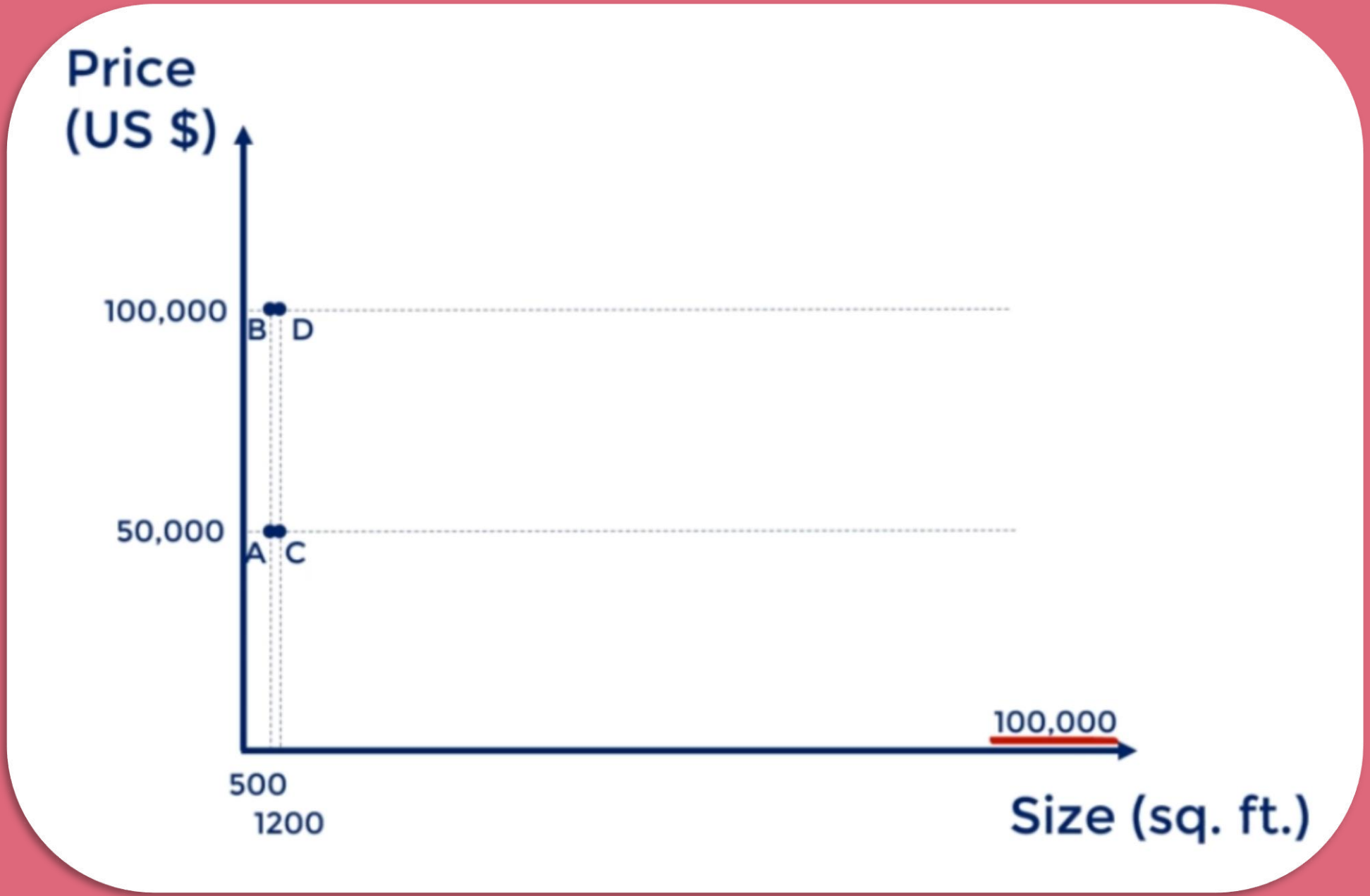

Price
(US $)
100,000
B
D
50,000
A
C
-1
1
Size_std

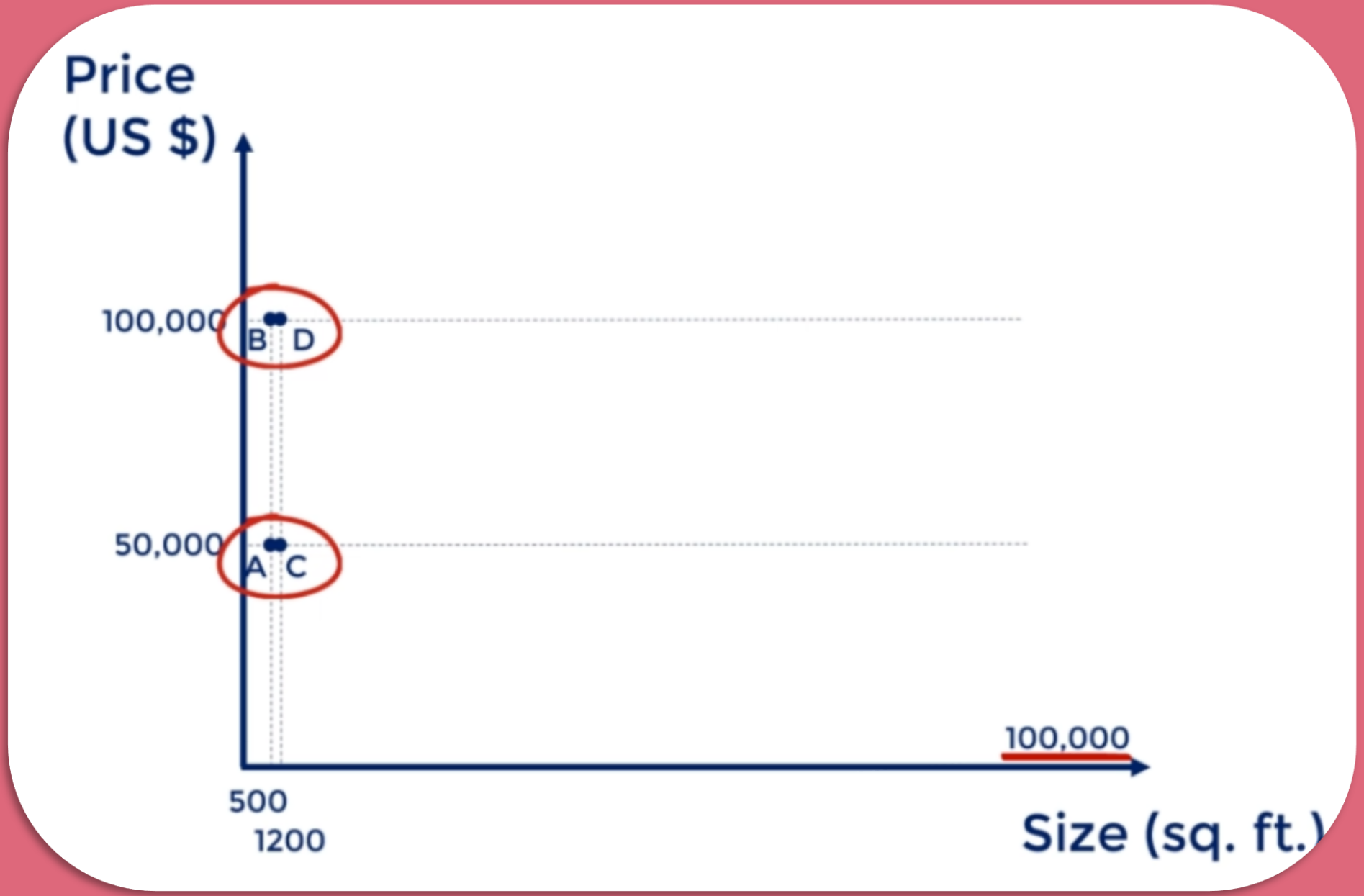




The ultimate aim of standardization is to reduce the weight of higher numbers

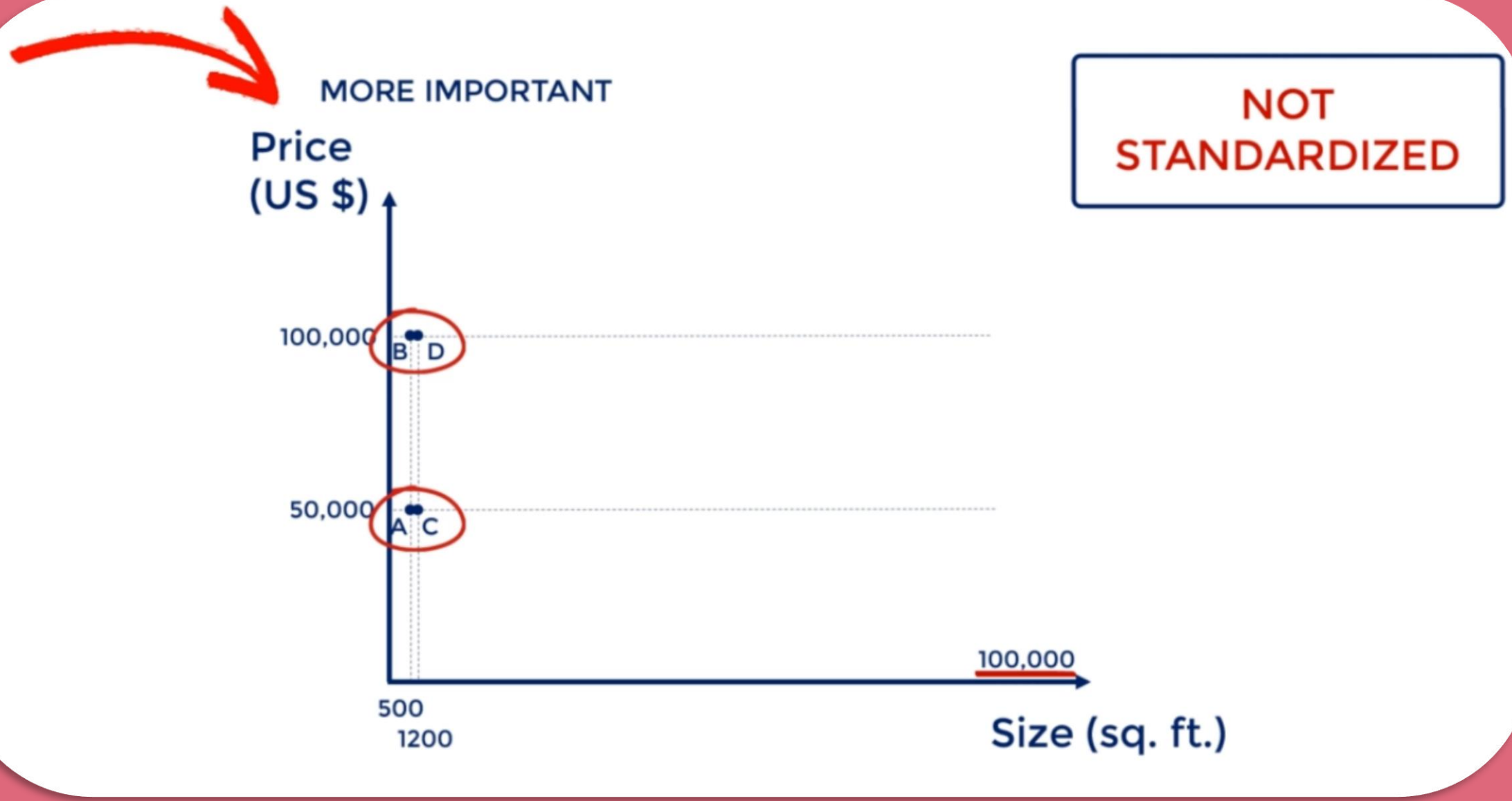


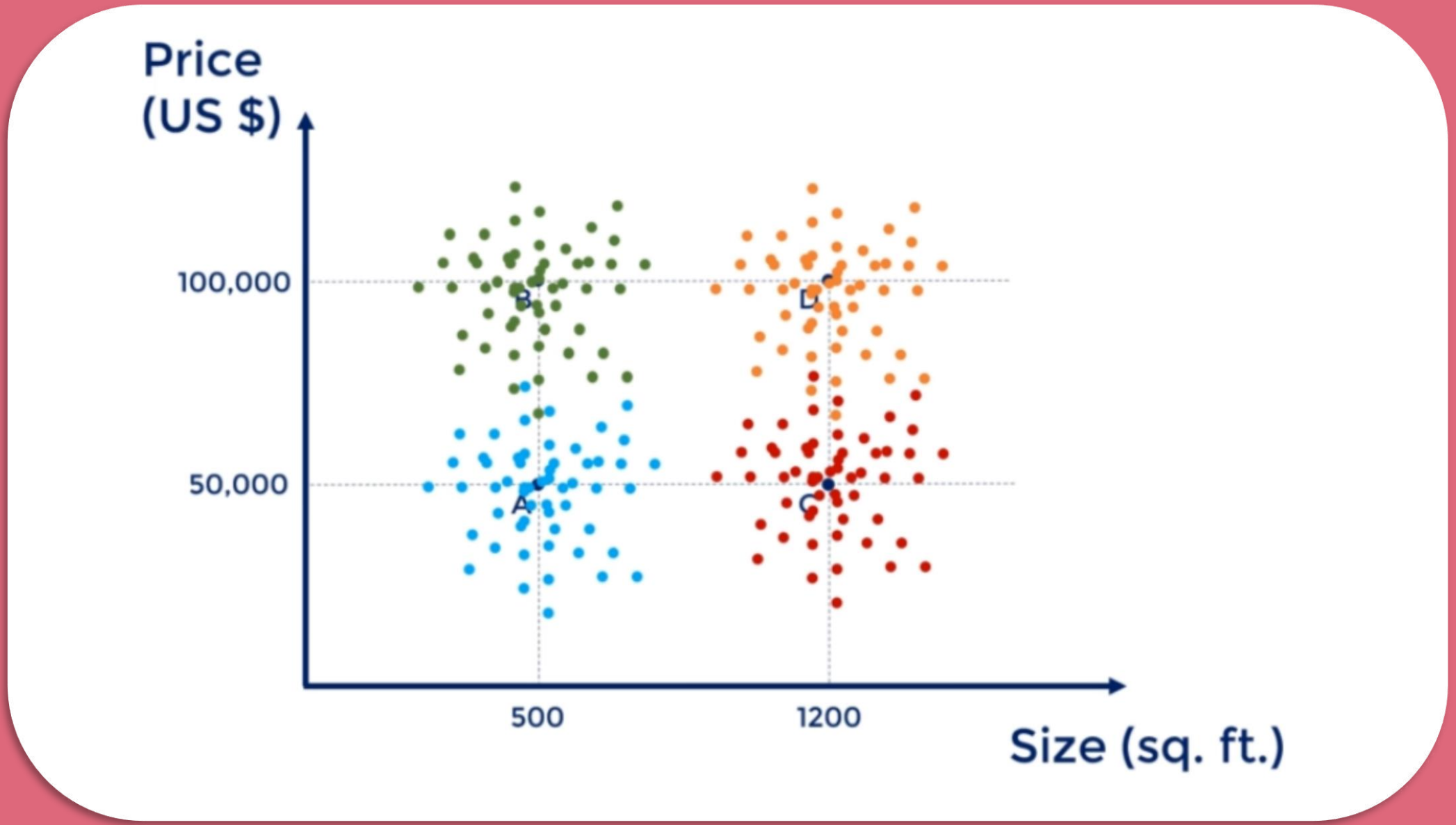

Price
(US $)
100,000
50,000
B
D
A
C
100,000
500
1200
Size (sq. ft.)

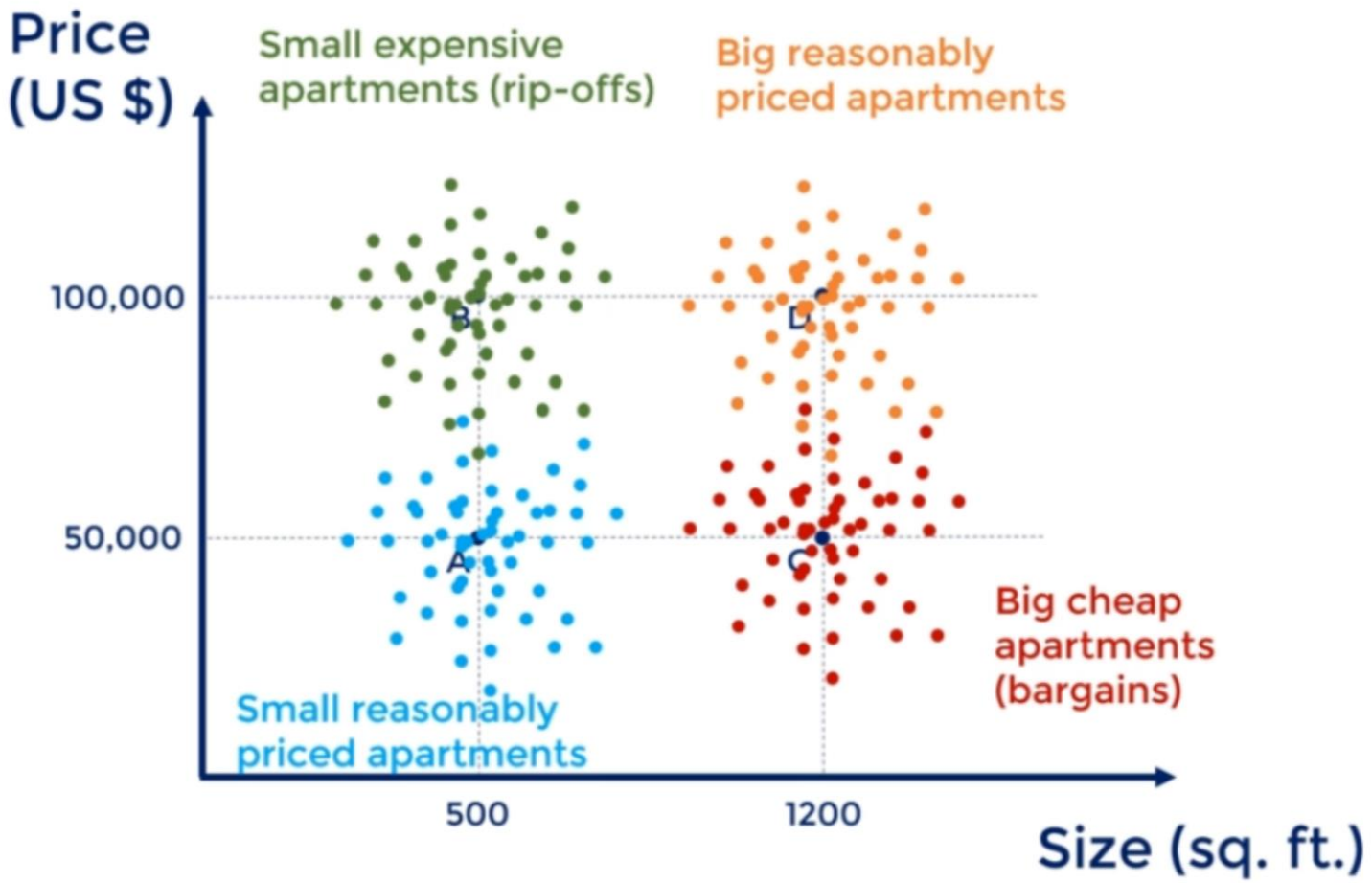

Price
(US $)
100,000
B
D
50,000
A
C
100,000
500
1200
Size (sq. ft.)

یک سؤال؟
کجا و کی استاندارد نکنیم؟

وقتی که دقیقا بدونیم یکی از ویژگیها از دیگری یا بقیه مهمتر هست. مثلاً در مثال قیمت و اندازه، قطعاً قیمت خیلی مهمتر هست. پس استاندارد نکنیم بهتره. اینجا تجربه بهتون کمک خواهد کرد که کم کم بفهمید ویژگی مهم کدوم هست


MORE IMPORTANT
NOT STANDARDIZED
Price (US $)
100,000
50,000
B
D
A
C
500
1200
100,000
Size (sq. ft.)


Price
(US $)
100,000
50,000
500
1200
Size (sq. ft.)
A
B
C
D



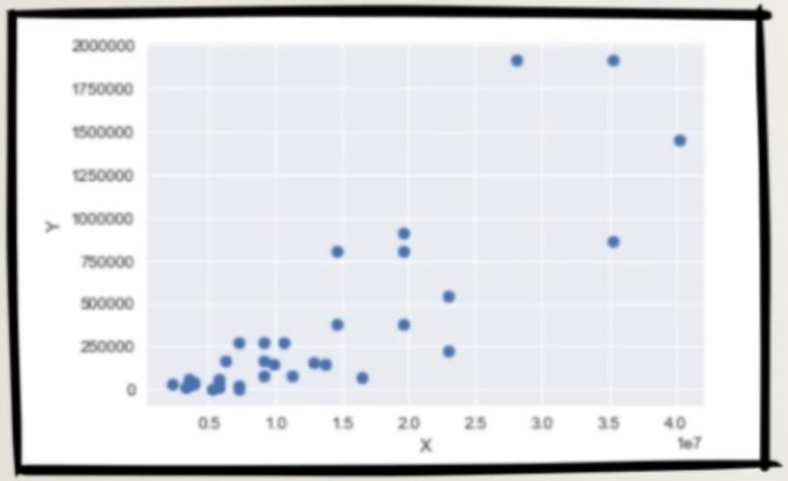
Clustering here can help us idenfity OVB

Types of analysis

- Exploratory
- Confirmatory
- Explanatory

- Get acquainted with the data
- Search for patterns
- Plan

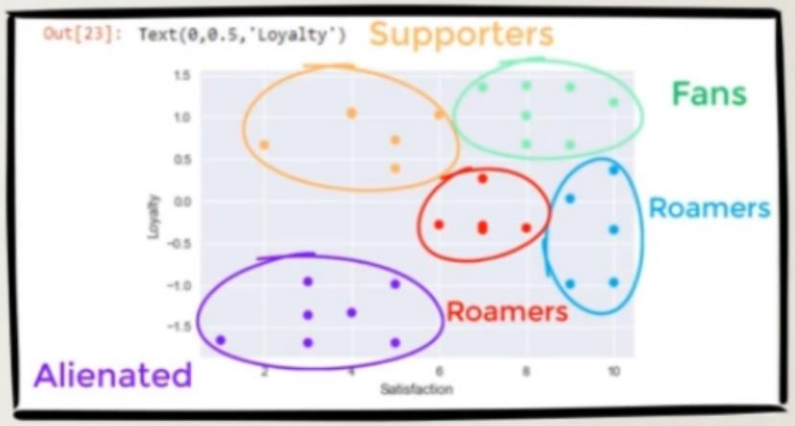
data visualization



descriptive

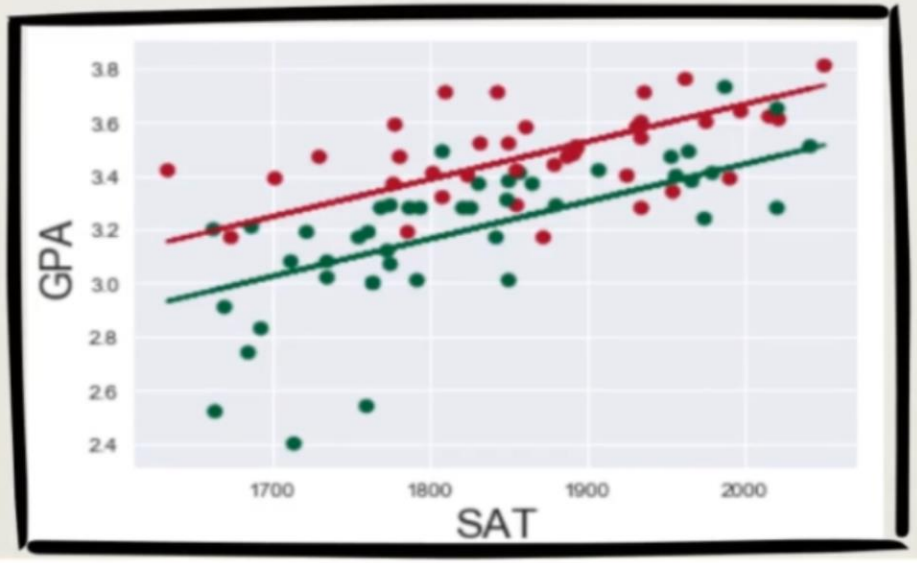
	SAT	GPA	Rand 1,2,3
count	84.000000	84.000000	84.000000
mean	1845.273810	3.330238	2.059524
std	104.530661	0.271617	0.855192
min	1634.000000	2.400000	1.000000
25%	1772.000000	3.190000	1.000000
50%	1846.000000	3.380000	2.000000
75%	1934.000000	3.502500	3.000000
max	2050.000000	3.810000	3.000000

clustering



CONFIRMATORY AND EXPLANATORY ANALYSIS

- Explain a phenomenon
- Confirm a hypothesis
- Validate previous research



Coding Time