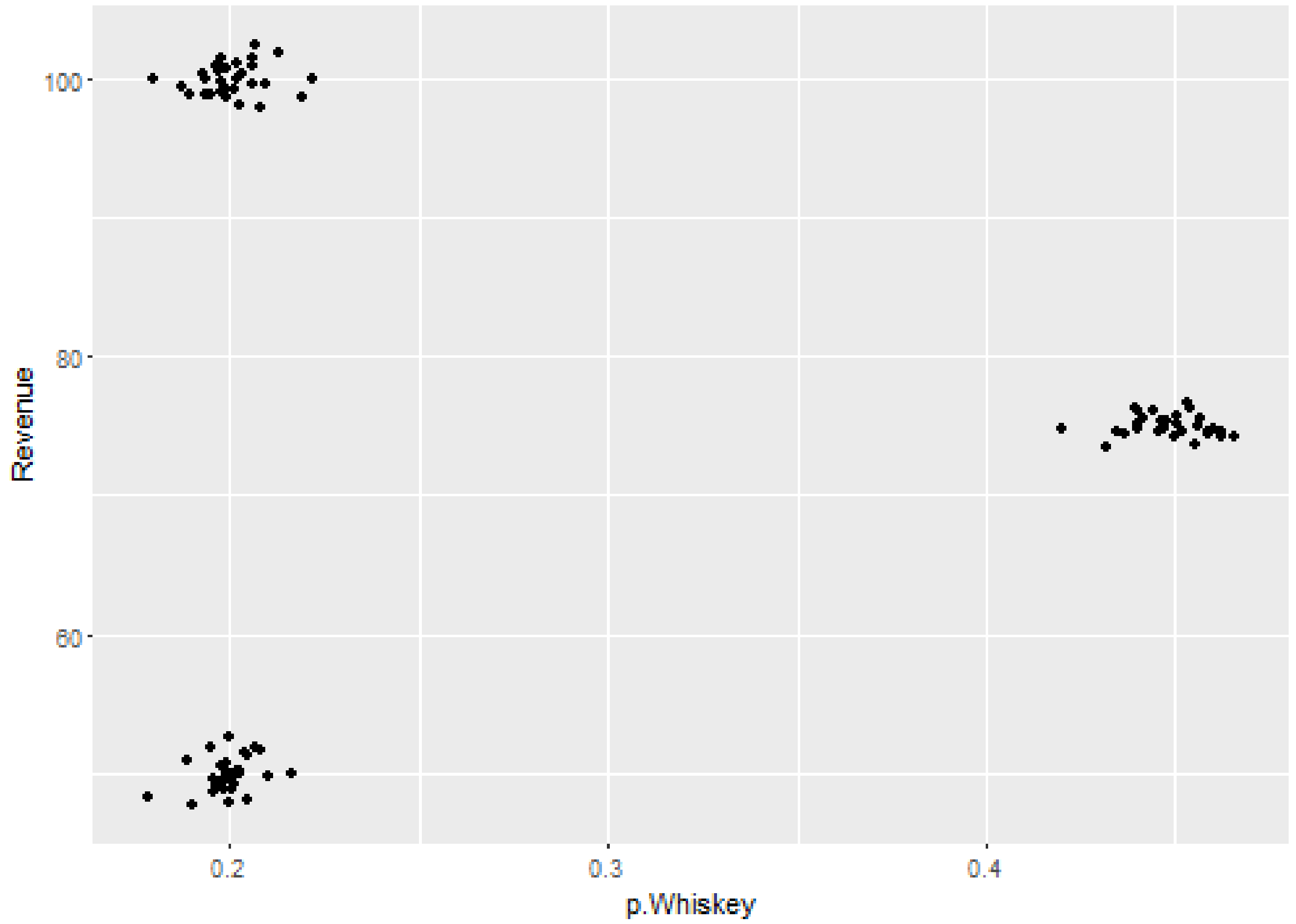
Clustering – Day 2

Clustering Kmeans: Good Value of K?

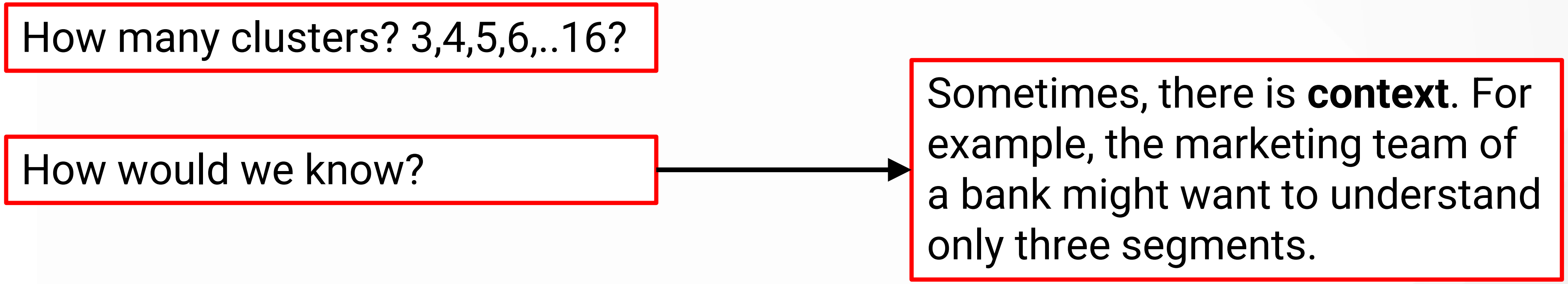


Store_id	P.Whisky	Revenue
1	0.40	80
2	0.20	60
3	0.35	40
4	0.22	90
5	0.45	75

How many clusters?
3, K=3

Clustering Kmeans: Good Value of K?

Income	Credit Limit	# Withdrawls	Card Usage	FICO	Age
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...



How many clusters? 3,4,5,6,..16?

How would we know?

Sometimes, there is **context**. For example, the marketing team of a bank might want to understand only three segments.

Clustering Kmeans: Good Value of K?

Income	Credit Limit	# Withdrawls	Card Usage	FICO	Age
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

How many clusters? 3,4,5,6,..16?

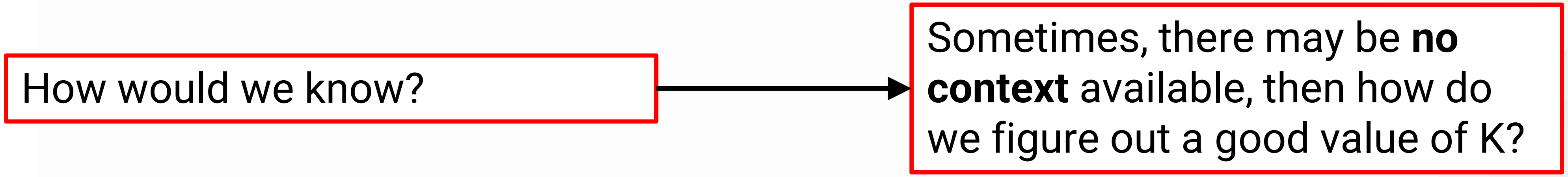
How would we know?

Sometimes, there may be **no context** available, then how do we figure out a good value of K?

Clustering Kmeans: Good Value of K?

Income	Credit Limit	# Withdrawls	Card Usage	FICO	Age
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

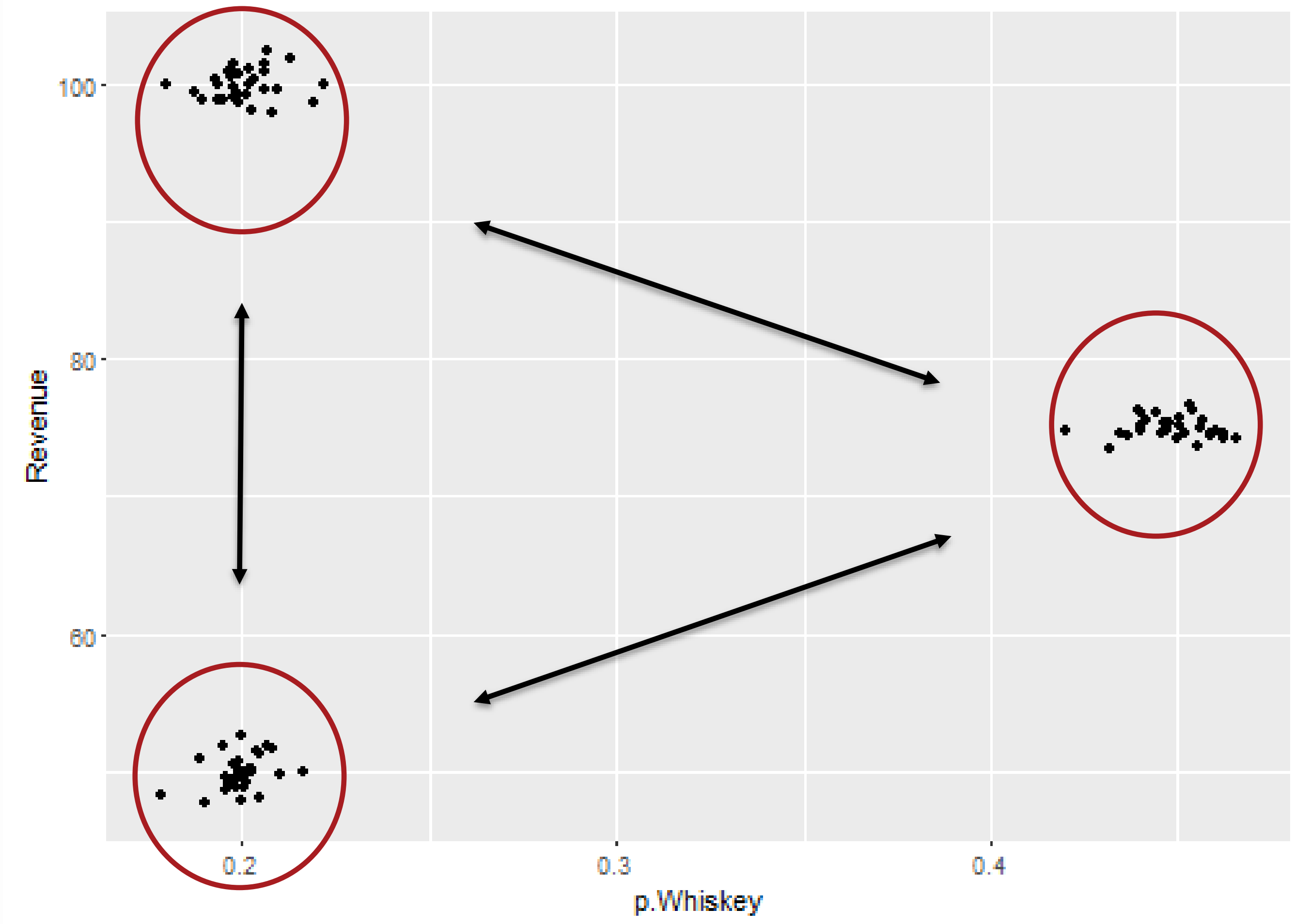
How many clusters? 3,4,5,6,..16?



How would we know?

Sometimes, there may be **no context** available, then how do we figure out a good value of K?

Clustering Kmeans: Good Value of K?



If we create 3 clusters:
i. Clusters are compact
ii. Clusters are far apart

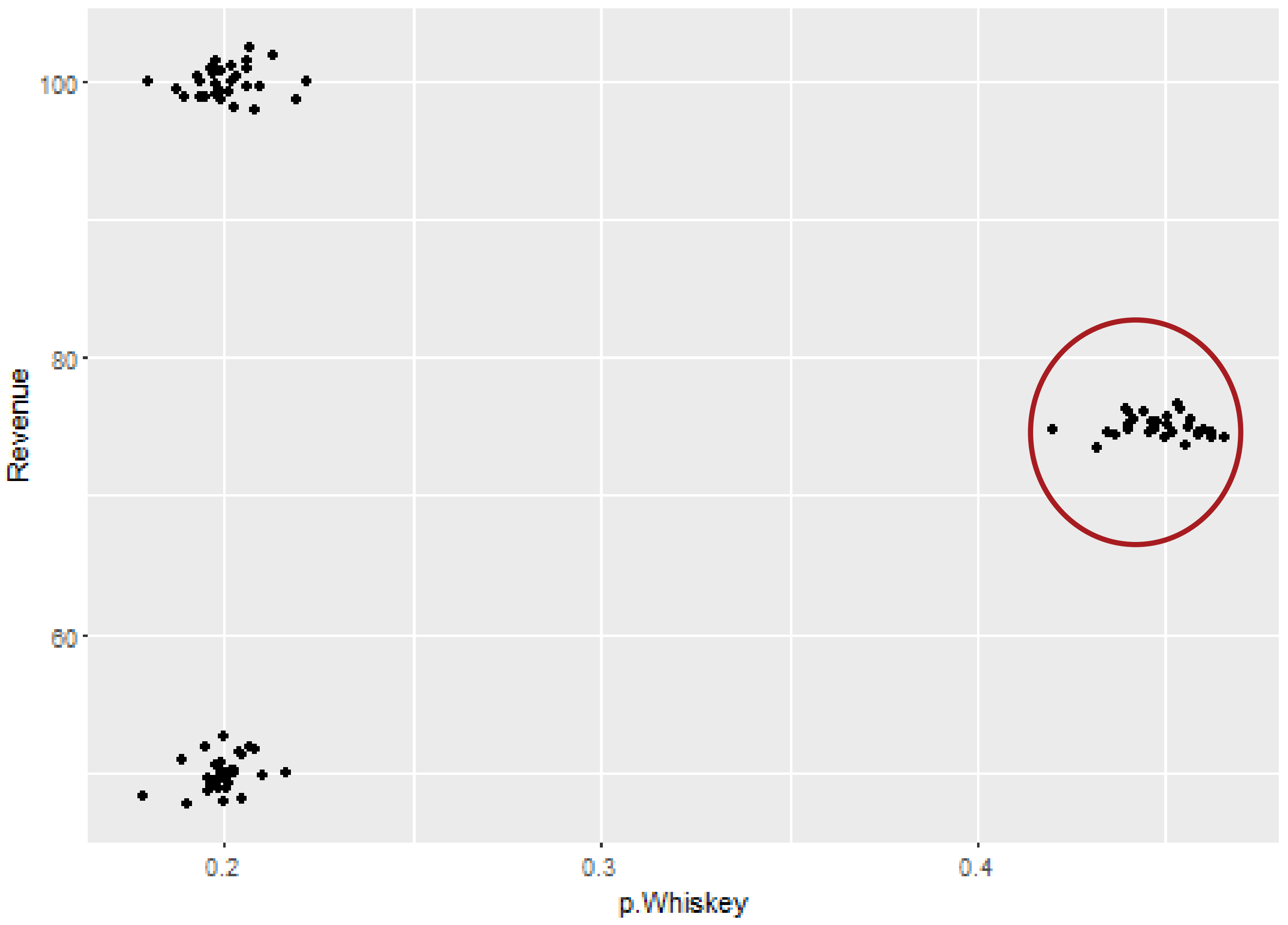
So, a good choice of K will lead to:
i. Compact Clusters
ii. Well separated clusters

Clustering Kmeans: Good Value of K?

#Clusters	Compactness
1	M1
2	M2
3	**M3**
4	M4
..	..

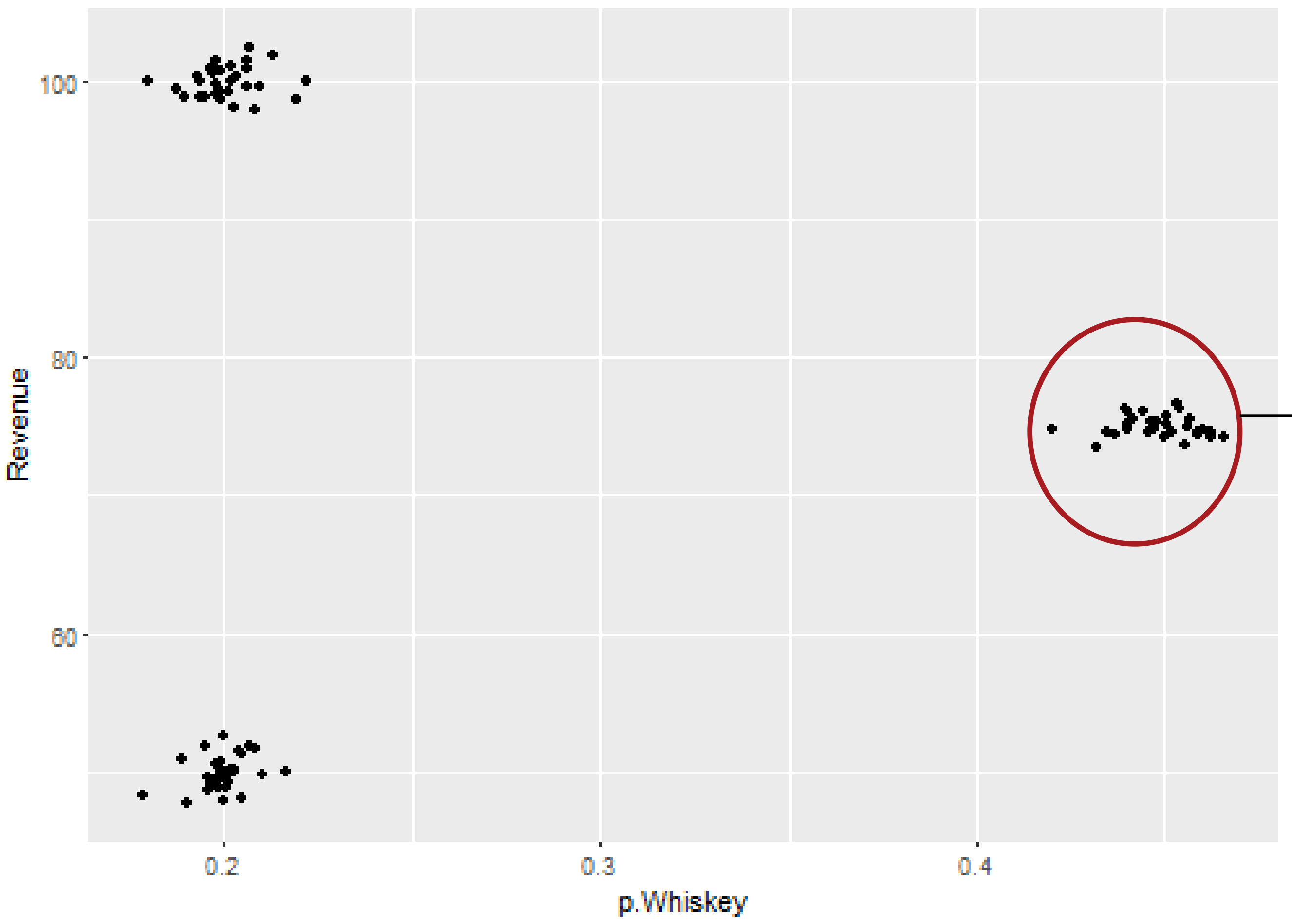
Choose K=3

Clustering Kmeans: Good Value of K?



Can we measure the compactness of clusters?

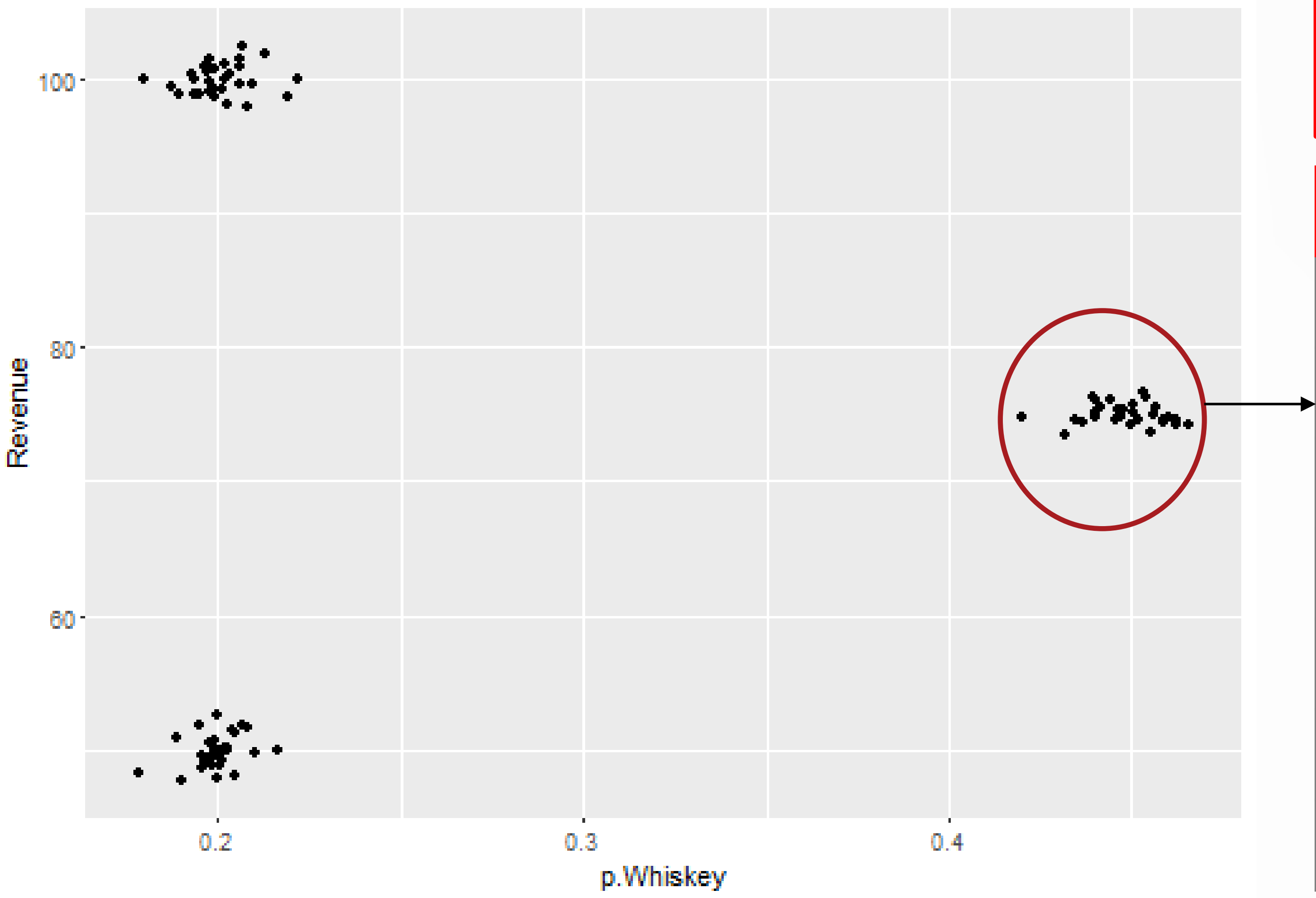
Clustering Kmeans: Good Value of K?



Can we measure the compactness of clusters?

Rev	P.Whisky
..	..
..	..
..	..
..	..
..	..
..	..
..	..

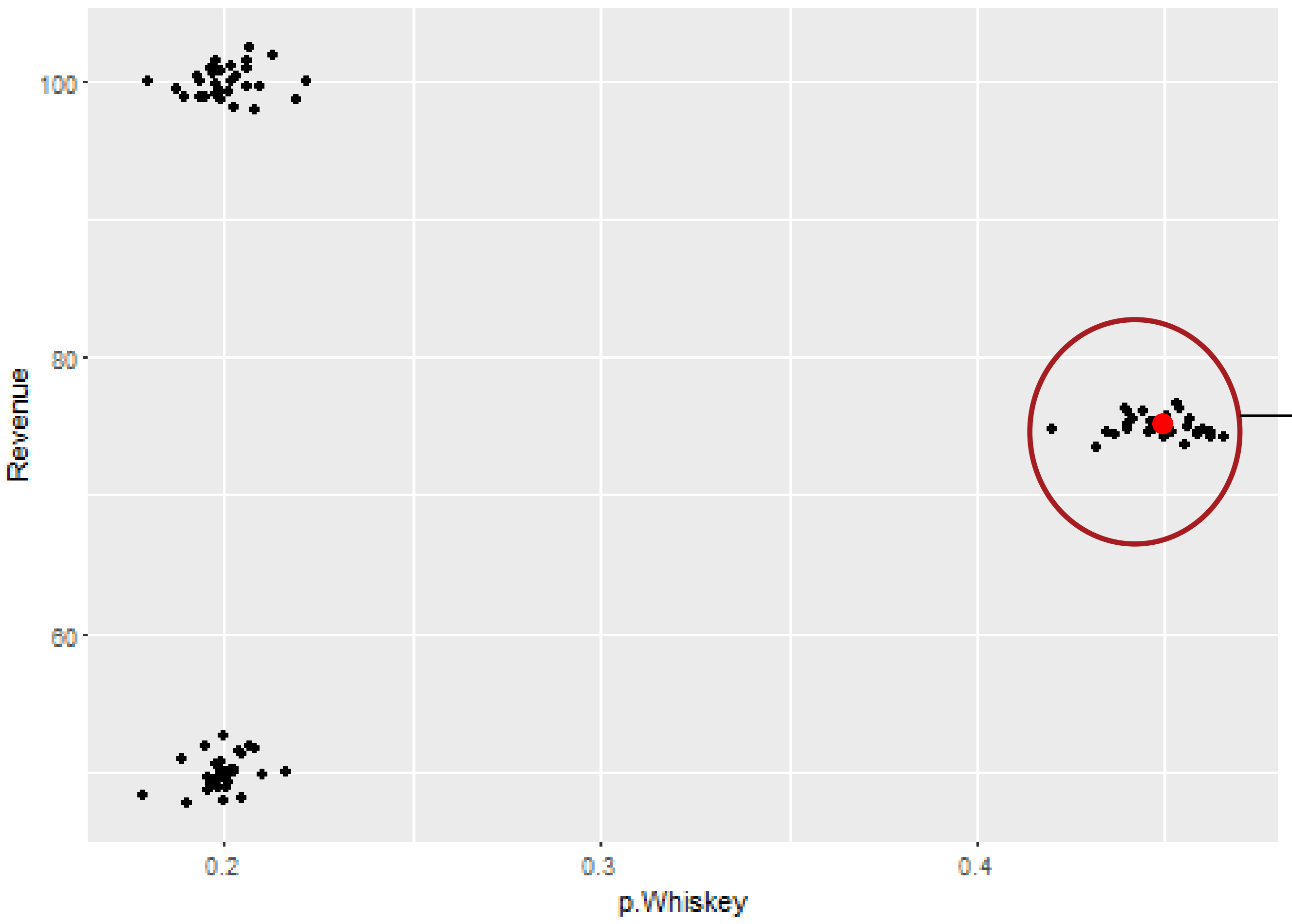
Clustering Kmeans: Good Value of K?



Can we measure the compactness of clusters?

Rev	P.Whisky
..	..
..	..
..	..
..	..
..	..
..	..
..	..

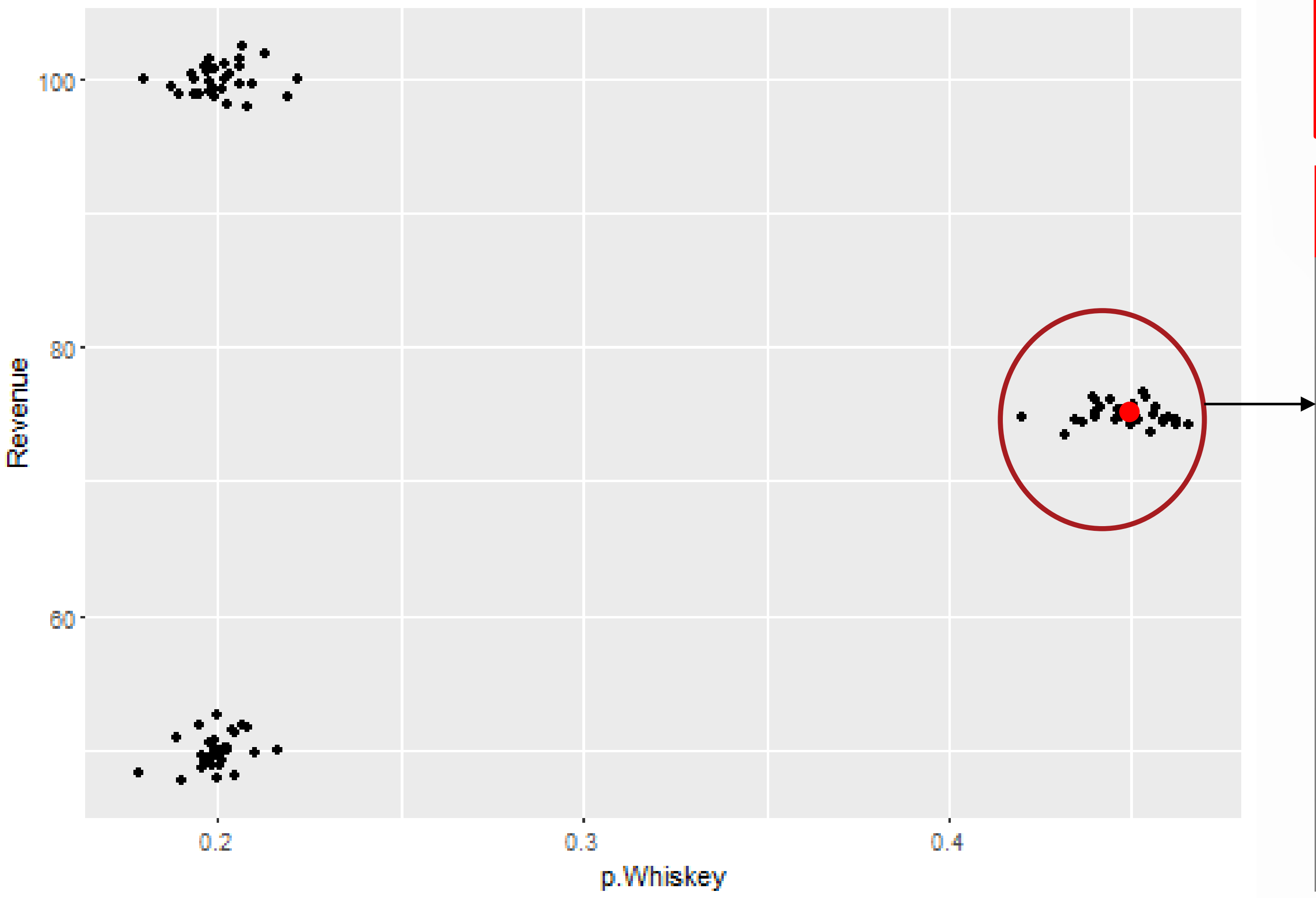
Clustering Kmeans: Good Value of K?



Can we measure the compactness of clusters?

Rev	P.Whisky
..	..
..	..
..	..
..	..
..	..
..	..
..	..

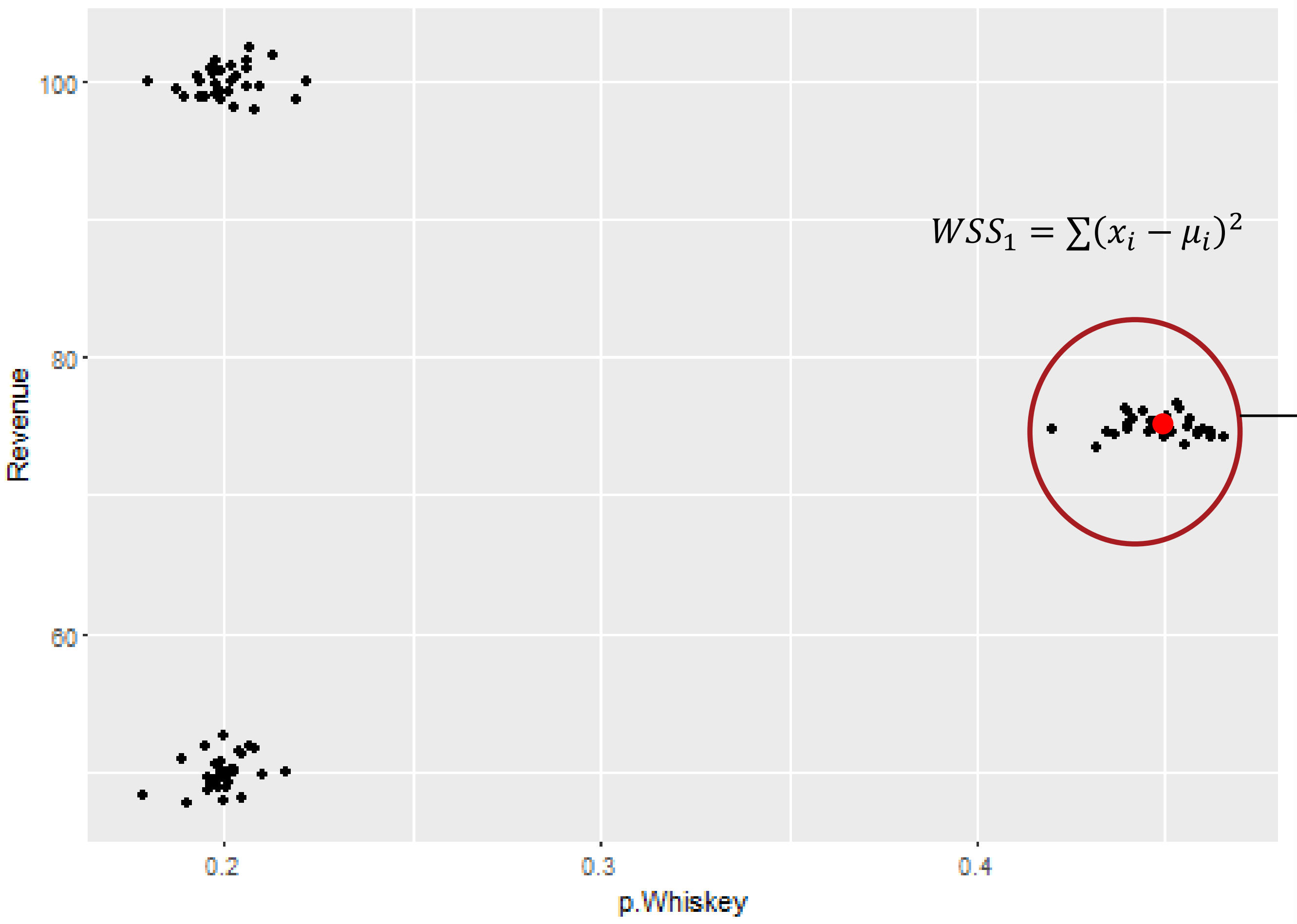
Clustering Kmeans: Good Value of K?



Can we measure the compactness of clusters?

Rev	P.Whisky
..	..
78	**0.42**
75	**0.45**
76	**0.42**
..	..
..	..
..	..

Clustering Kmeans: Good Value of K?

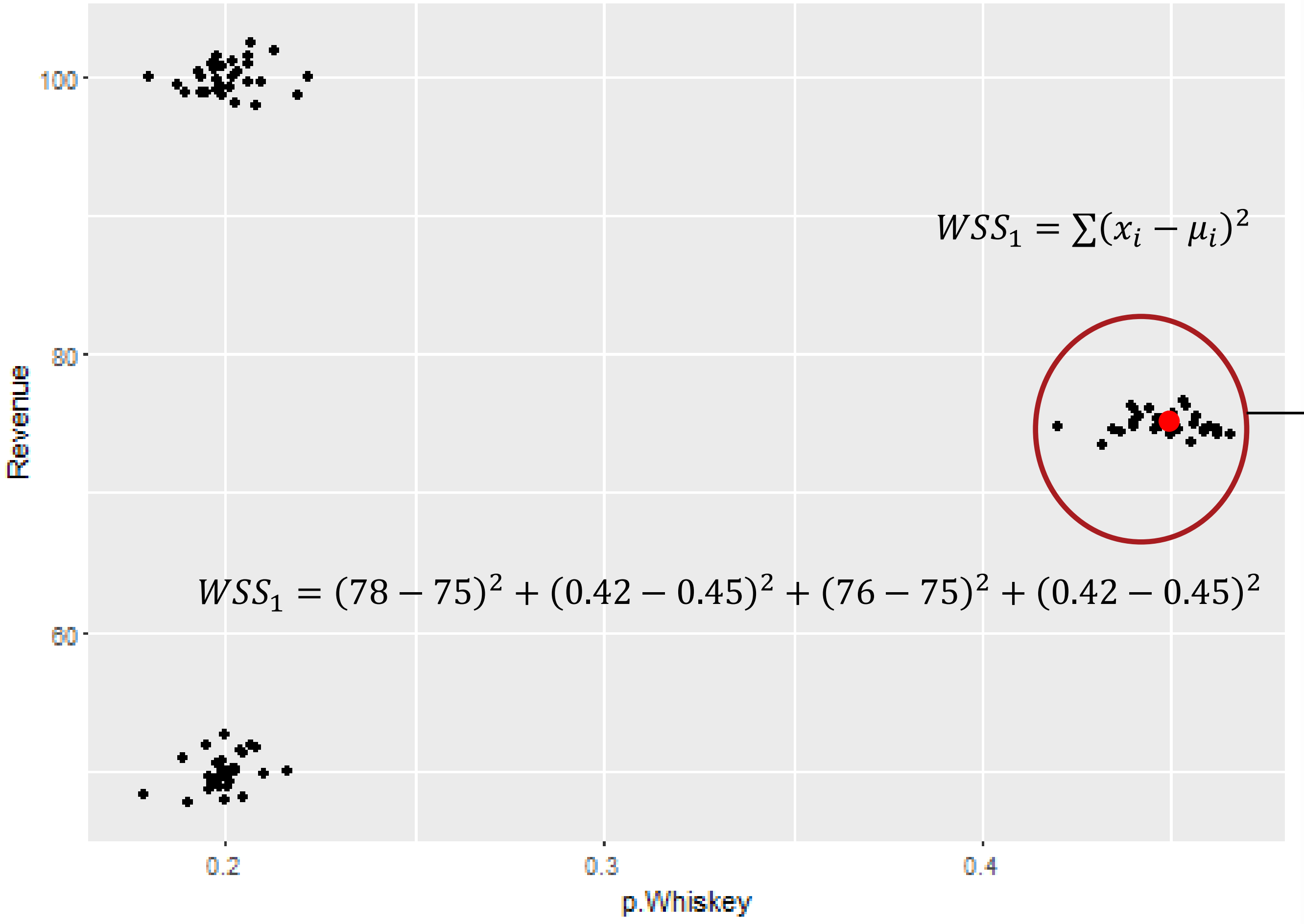


$$WSS_1 = \sum(x_i - \mu_i)^2$$

Can we measure the compactness of clusters?

Rev	P.Whisky
..	..
78	0.42
75	0.45
76	0.42
..	..
..	..
..	..

Clustering Kmeans: Good Value of K?



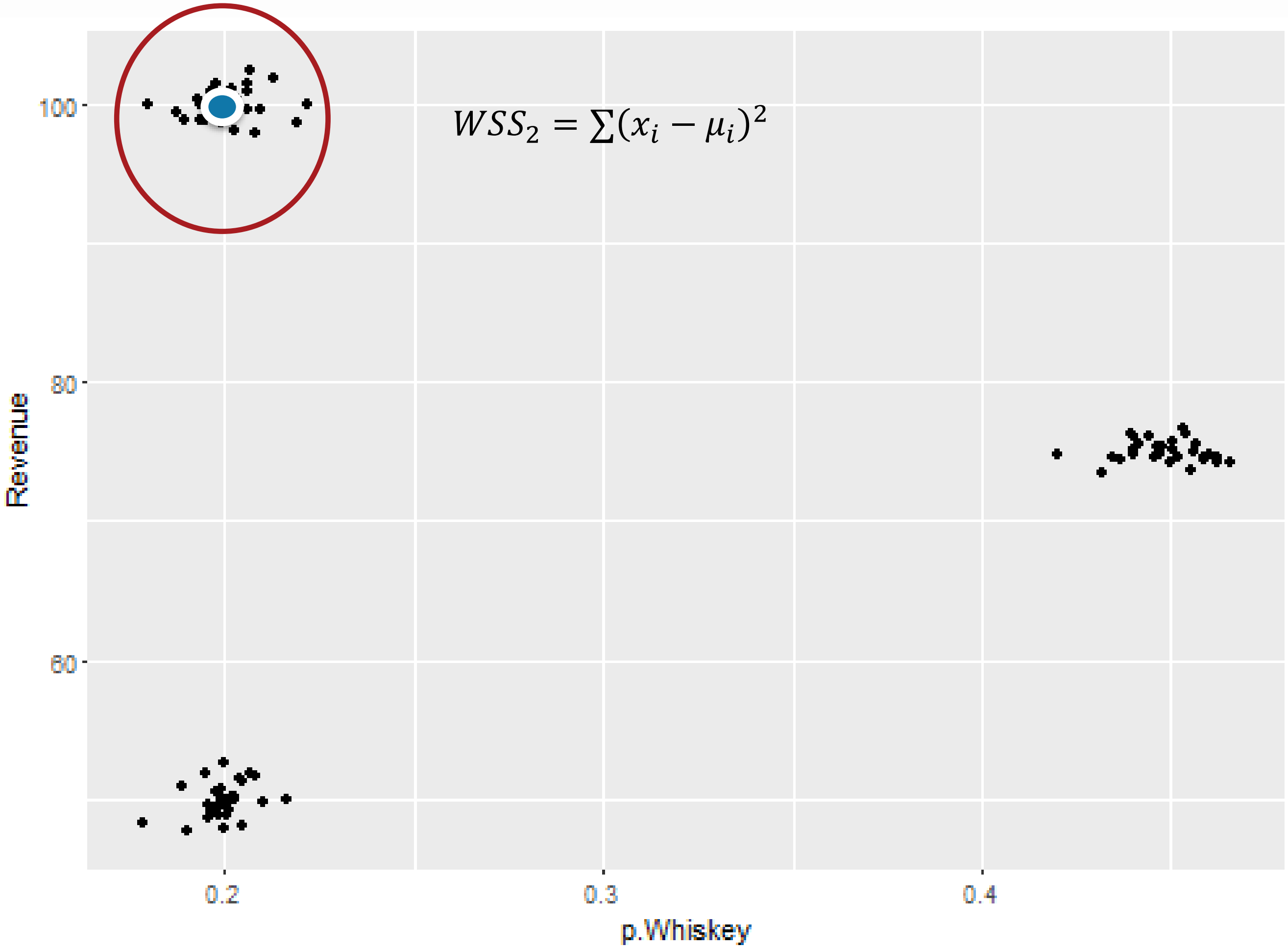
$$WSS_1 = \sum (x_i - \mu_i)^2$$

$$WSS_1 = (78 - 75)^2 + (0.42 - 0.45)^2 + (76 - 75)^2 + (0.42 - 0.45)^2$$

Can we measure the compactness of clusters?

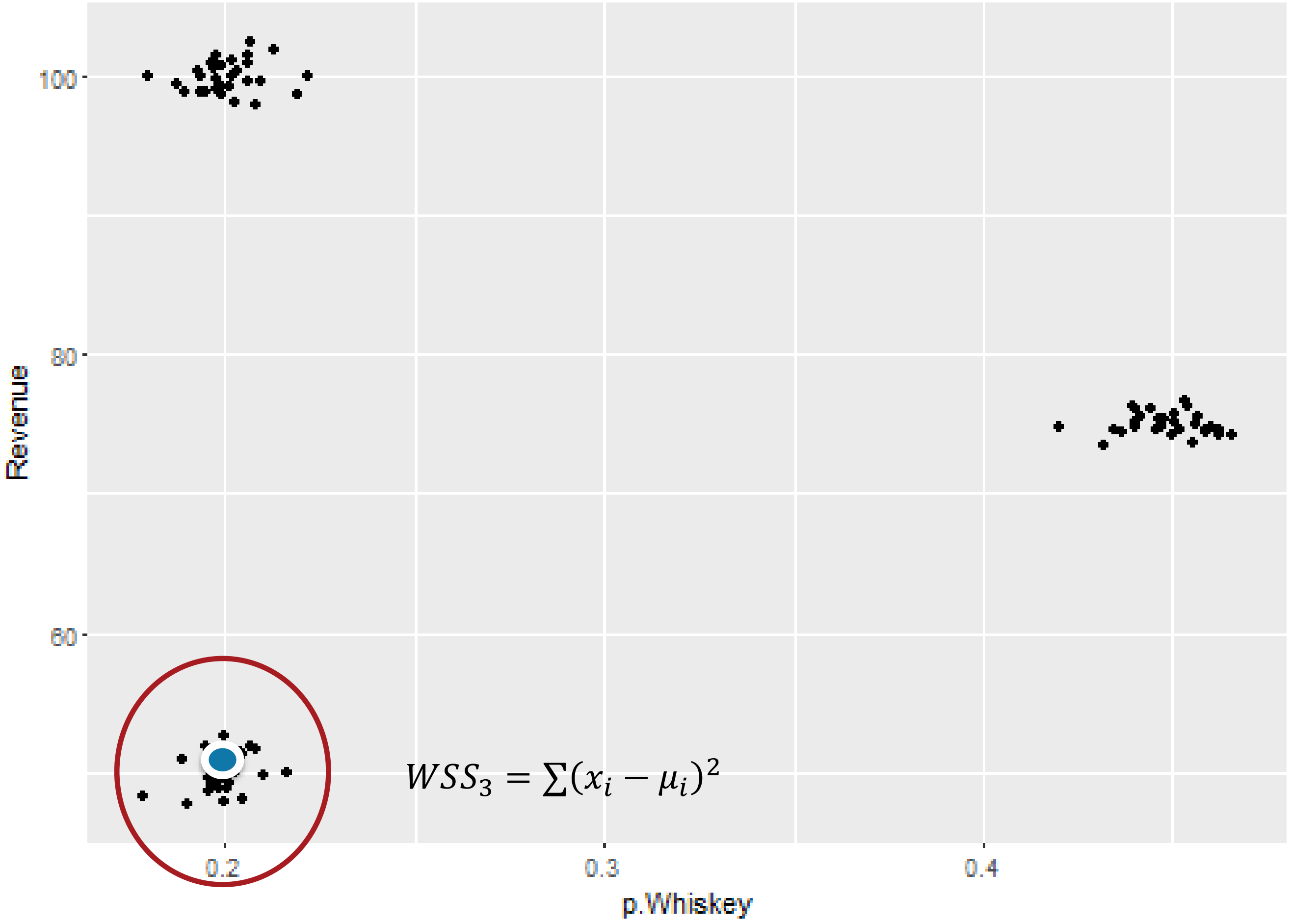
Rev	P.Whisky
..	..
78	**0.42**
75	**0.45**
76	**0.42**
..	..
..	..
..	..

Clustering Kmeans: Good Value of K?



Can we measure the compactness of clusters?

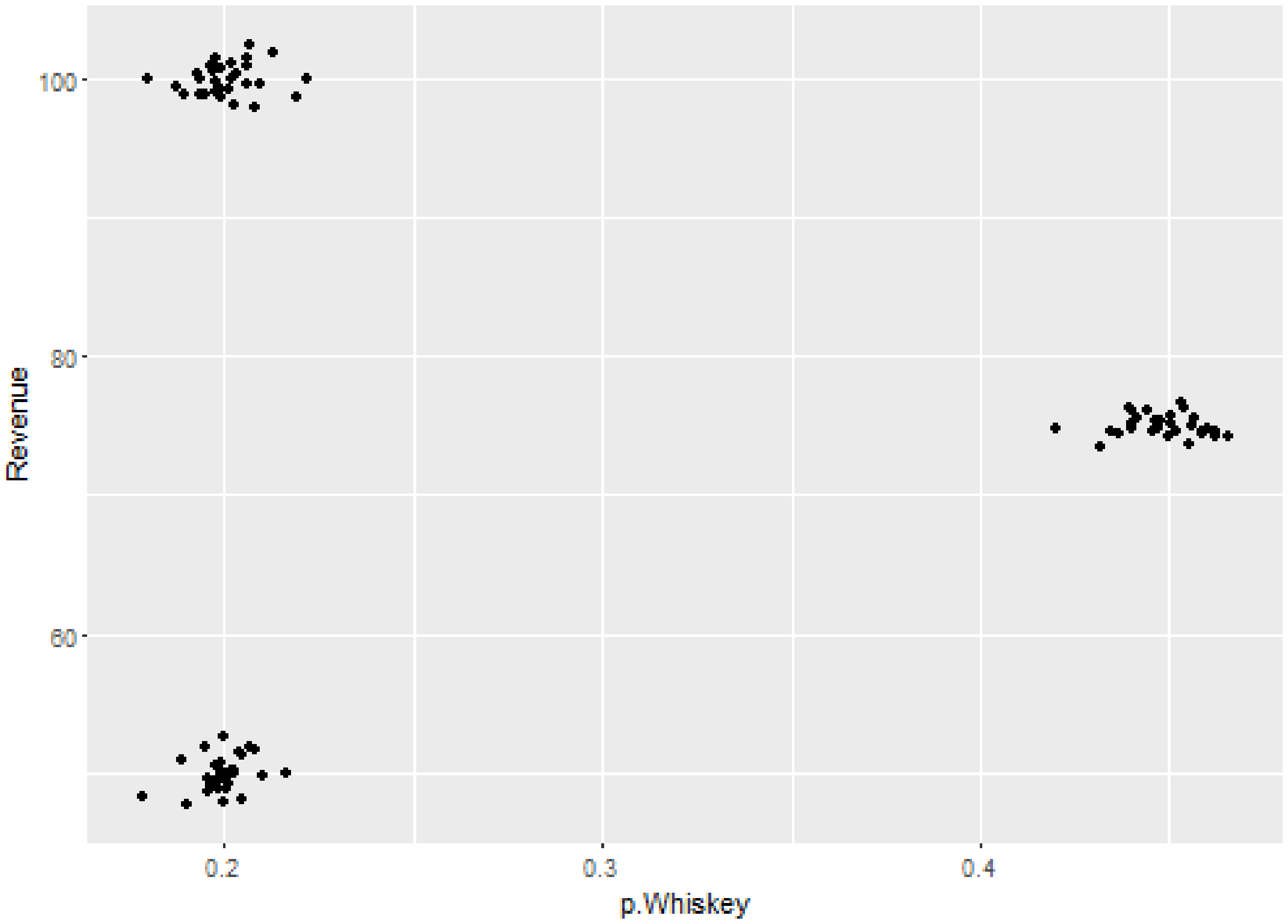
Clustering Kmeans: Good Value of K?



$$WSS_3 = \sum(x_i - \mu_i)^2$$

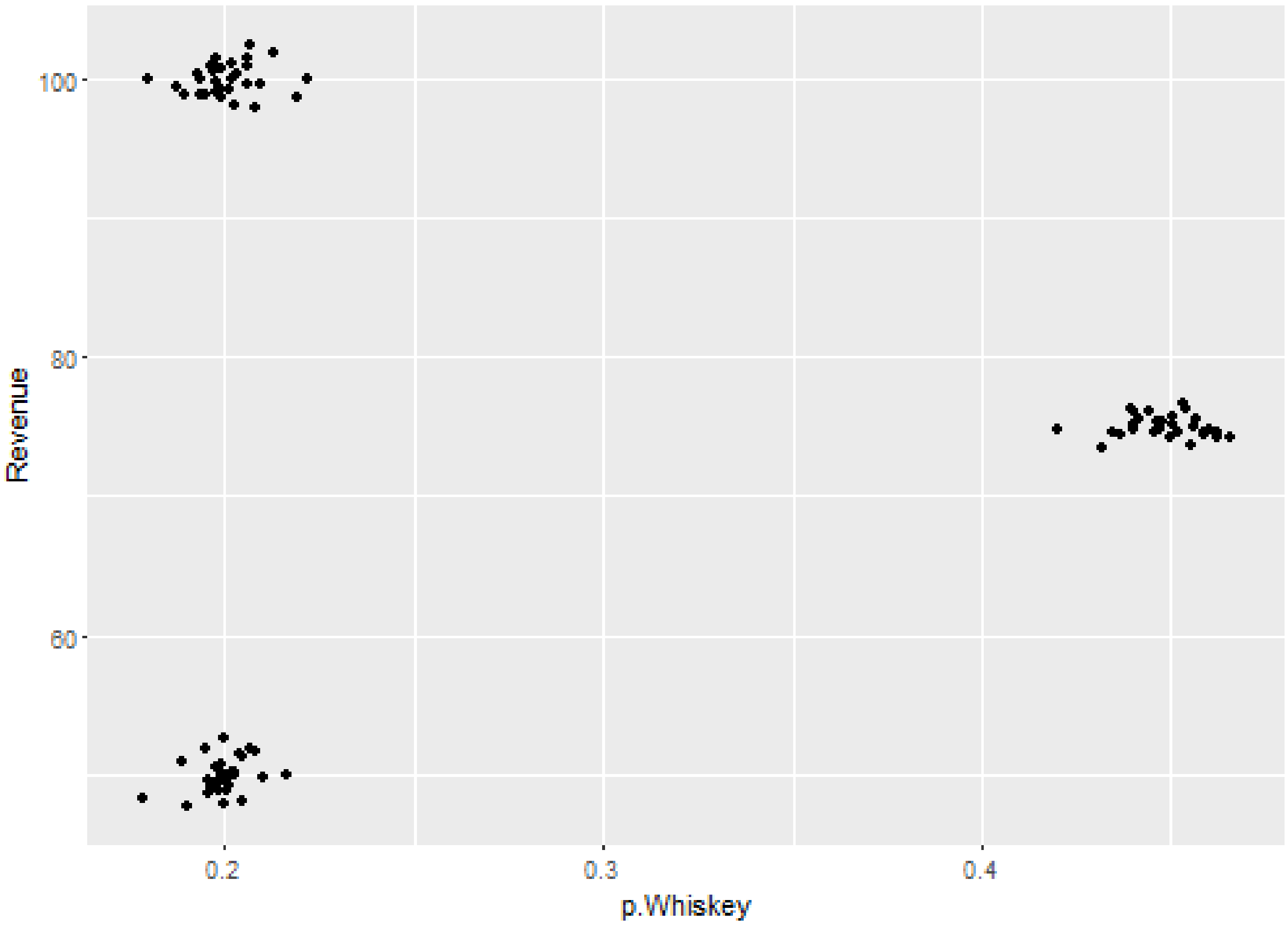
Can we measure the compactness of clusters?

Clustering Kmeans: Good Value of K?



- Can we measure the compactness of clusters?
- $WSS_{Total} = WSS_1 + WSS_2 + WSS_3$

Clustering Kmeans: Good Value of K?

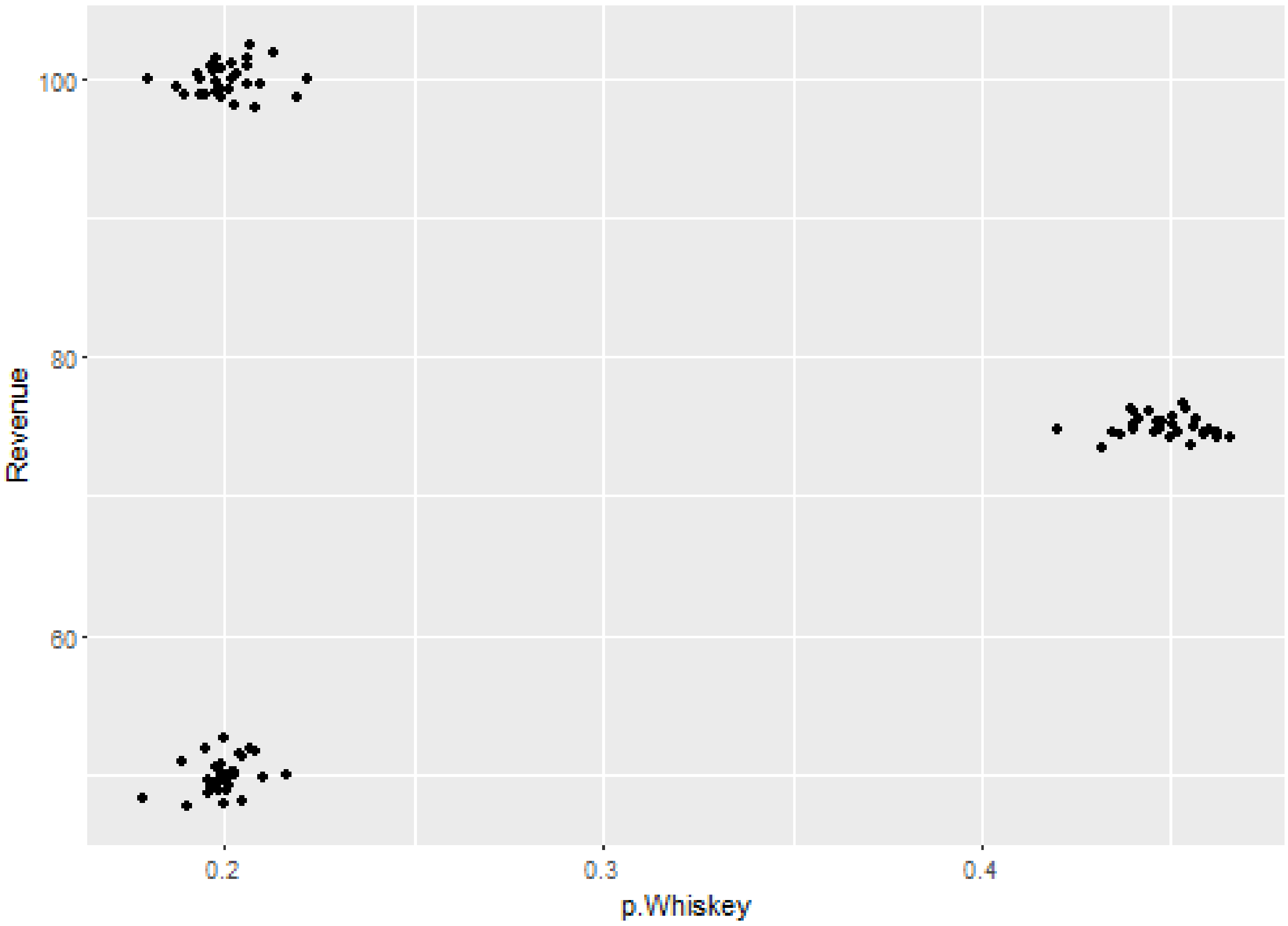


- Can we measure the compactness of clusters?
- $WSS_{Total} = WSS_1 + WSS_2 + WSS_3$

or

- $WSS_{Average} = \frac{1}{3}(WSS_1 + WSS_2 + WSS_3)$

Clustering Kmeans: Good Value of K?



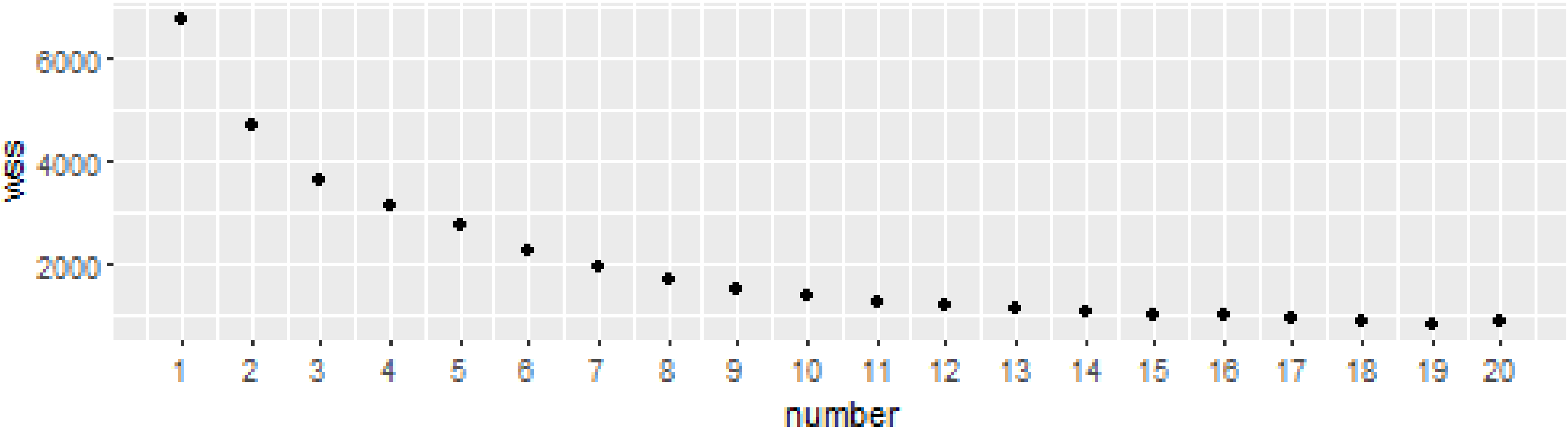
- Can we measure the compactness of clusters?
- $WSS_{Total} = WSS_1 + WSS_2 + WSS_3$

or

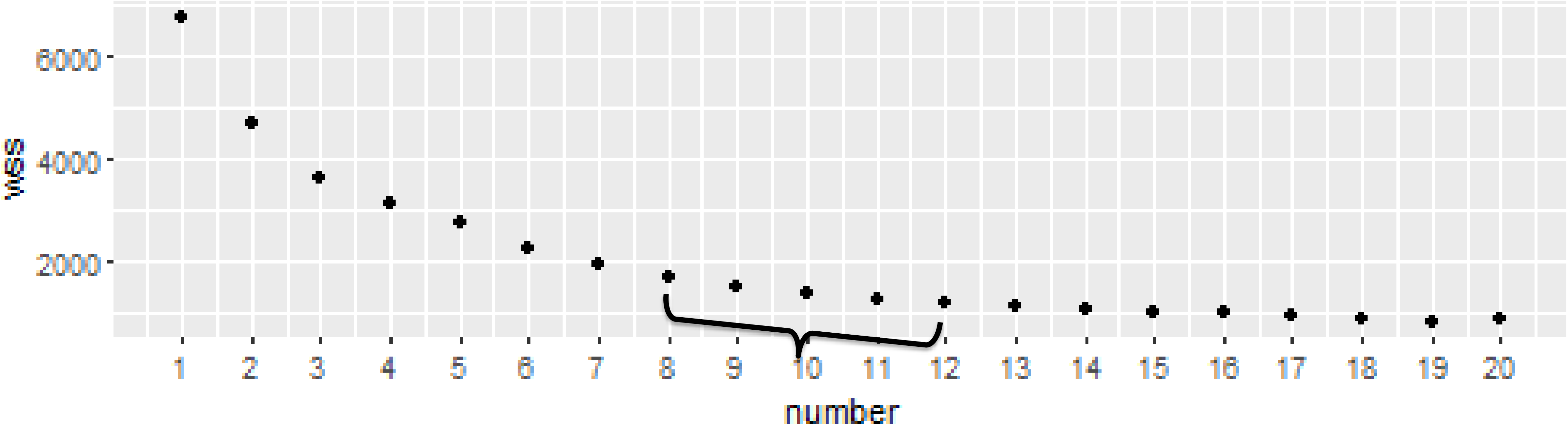
- $WSS_{Average} = \frac{1}{3}(WSS_1 + WSS_2 + WSS_3)$

#Clusters	WSS
1	M1
2	M2
3	M3
4	M4
..	..

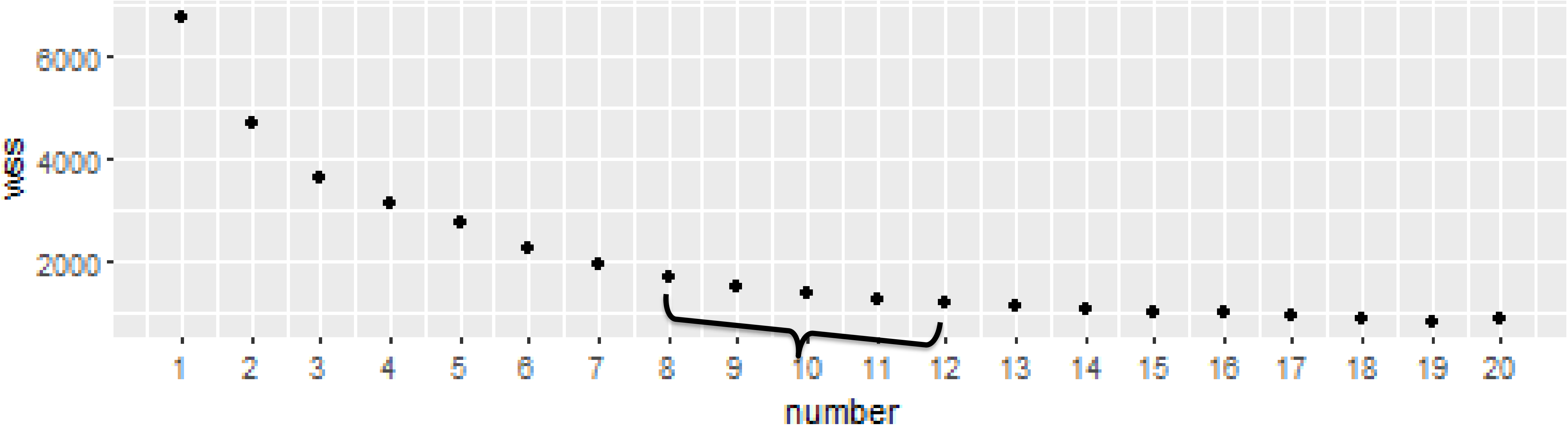
Clustering Kmeans: Good Value of K?



Clustering Kmeans: Good Value of K?



Clustering Kmeans: Good Value of K?



Anything between **8 to 12** clusters is a good number.

Clustering Class Exercise

Age	Weight	Cluster
30	67	C1
35	70	C1
25	64	C2
23	62	C2

Center	X (Age)	Y (Weight)
C1	32	68
C2	24	63

Find out $WSS_{average}$ for the data given above.

Clustering Class Exercise

Age	Weight	Cluster
30	67	C1
35	70	C1
25	64	C2
23	62	C2

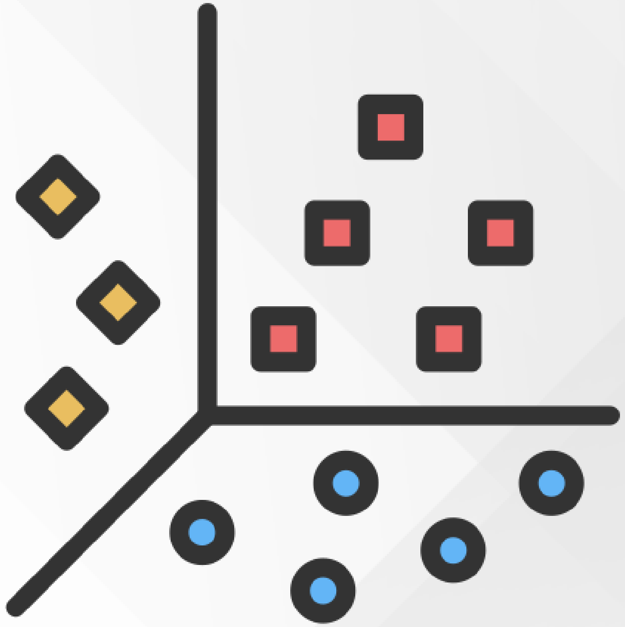
Center	X (Age)	Y (Weight)
C1	32	68
C2	24	63

Find out $WSS_{average}$ for the data given above.

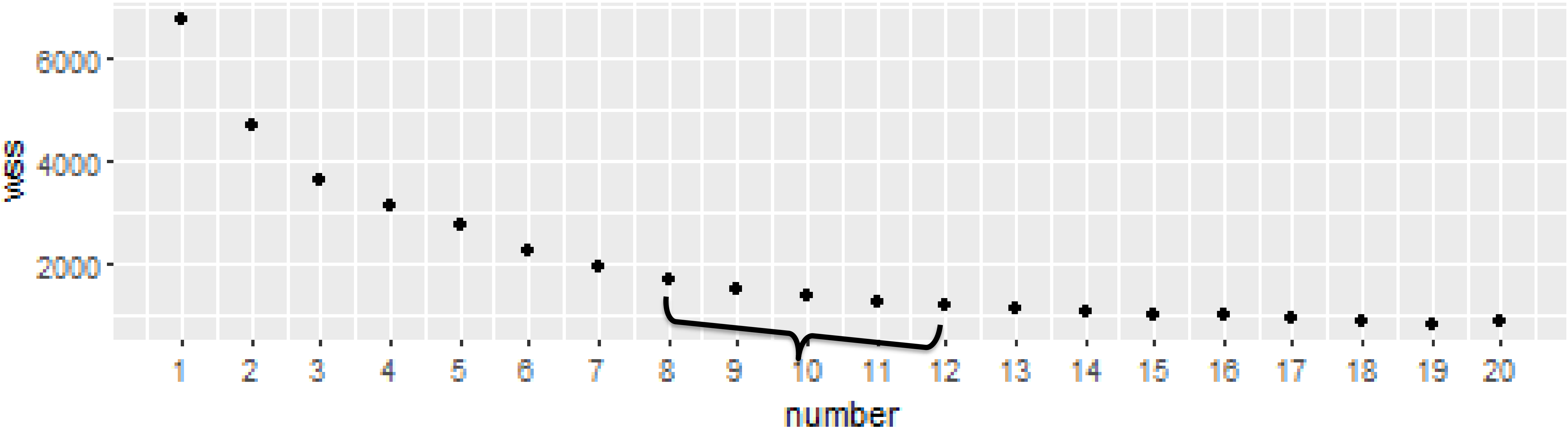
See **Numerical Example Clustering.xlsx** in sheet Class Exercise WSS

Clustering: Elbow Curve Demo

Links: https://github.com/Gunnvant/Self-Paced-Content/tree/main/python/clustering/Class%20Demos



Clustering Kmeans: Good Value of K?

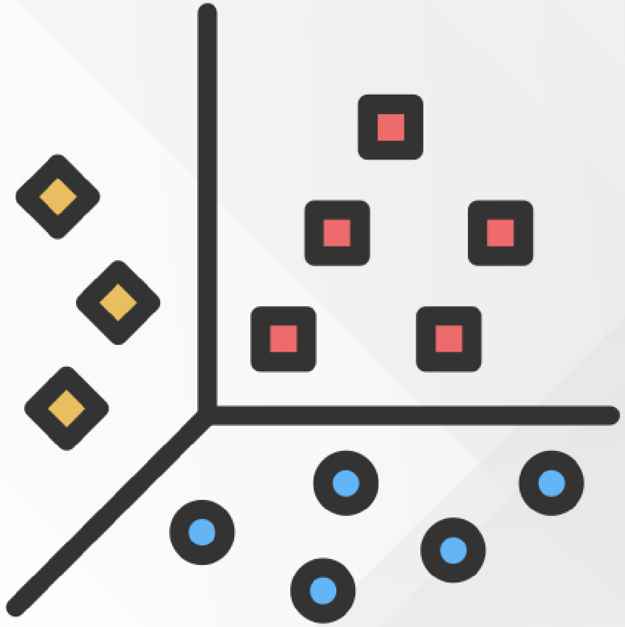


Anything between **8 to 12** clusters is a good number.

Now what?

Clustering: Profiling Clusters

- Once you finalize the number of clusters, you will then need to describe these clusters.
- This is done via a process known as **cluster profiling.**



Clustering Kmeans: Good Value of K? Cluster Profiles

Revenue	P.Whisky	Cluster
..	..	1
..	..	1
..	..	2
..	..	2
..	..	3
..	..	3
..	..	1

Mean Revenue= 120, std = 10

Mean Revenue Cluster 1 = 200

Mean Revenue Cluster 2 = 125

Mean Revenue Cluster 3 = 75

Mean P.Whisky= 0.20, std= 0.10

Mean P.Whisky Cluster 1 = 0.40

Mean P.Whisky Cluster 2 = 0.21

Mean P.Whisky Cluster 3 = 0.05

Clustering Kmeans: Good Value of K? Cluster Profiles

Revenue	P.Whisky	Cluster
..	..	1
..	..	1
..	..	2
..	..	2
..	..	3
..	..	3
..	..	1

Mean Revenue= 120, std = 10

Mean Revenue Cluster 1 = 200

Mean Revenue Cluster 2 = 125

Mean Revenue Cluster 3 = 75

Revenue in cluster 1 is more than average, so cluster 1 is a cluster with high revenue.

Mean P.Whisky= 0.20, std= 0.10

Mean P.Whisky Cluster 1 = 0.40

Mean P.Whisky Cluster 2 = 0.21

Mean P.Whisky Cluster 3 = 0.05

Clustering Kmeans: Good Value of K? Cluster Profiles

Revenue	P.Whisky	Cluster
..	..	1
..	..	1
..	..	2
..	..	2
..	..	3
..	..	3
..	..	1

Mean Revenue= 120, std = 10

Mean Revenue Cluster 1 = 200

Mean Revenue Cluster 2 = 125

Mean Revenue Cluster 3 = 75

Revenue in cluster 3 is less than average, so cluster 1 is a cluster with low revenue.

Mean P.Whisky= 0.20, std= 0.10

Mean P.Whisky Cluster 1 = 0.40

Mean P.Whisky Cluster 2 = 0.21

Mean P.Whisky Cluster 3 = 0.05

Clustering Kmeans: Good Value of K? Cluster Profiles

Revenue	P.Whisky	Cluster
..	..	1
..	..	1
..	..	2
..	..	2
..	..	3
..	..	3
..	..	1

Mean Revenue = 120, std = 10

Mean Revenue Cluster 1 = 200

Mean Revenue Cluster 2 = 125

Mean Revenue Cluster 3 = 75

Mean P.Whisky = 0.20, std = 0.10

Mean P.Whisky Cluster 1 = 0.40

Mean P.Whisky Cluster 2 = 0.21

Mean P.Whisky Cluster 3 = 0.05

Mean Revenue= 120, std = 10

Z Revenue Cluster 1 = 8

Z Revenue Cluster 2 = 0.5

Z Revenue Cluster 3 = -4.5

Mean P.Whisky= 0.20, std= 0.10

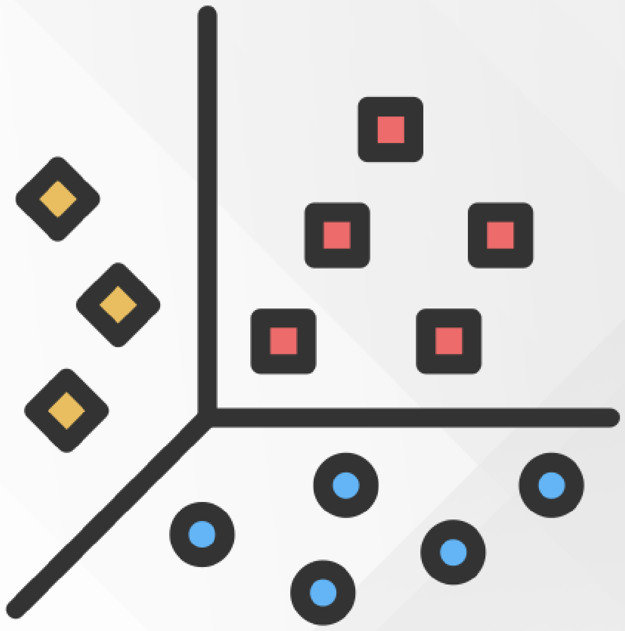
Z P.Whisky Cluster 1 = 2

Z P.Whisky Cluster 2 = 1

Z P.Whisky Cluster 3 = -1.5

Cluster Profiling: Code Demo

Links: https://github.com/Gunnvant/Self-Paced-Content/tree/main/python/clustering/Class%20Demos



Thank You!