Clustering – Day 1

Learning Outcomes

On completion of the session, you will be able to:

- Identify business scenarios in which clustering can be helpful
- Describe in pseudo code how k-means clustering algorithm works
- Perform necessary data transformations in order to carry out k-means
- Select optimal number of clusters
- Analyze clusters in the business context and generate segments
- Given a business problem and a dataset, build a cluster model
- Create business relevant segments using a cluster model

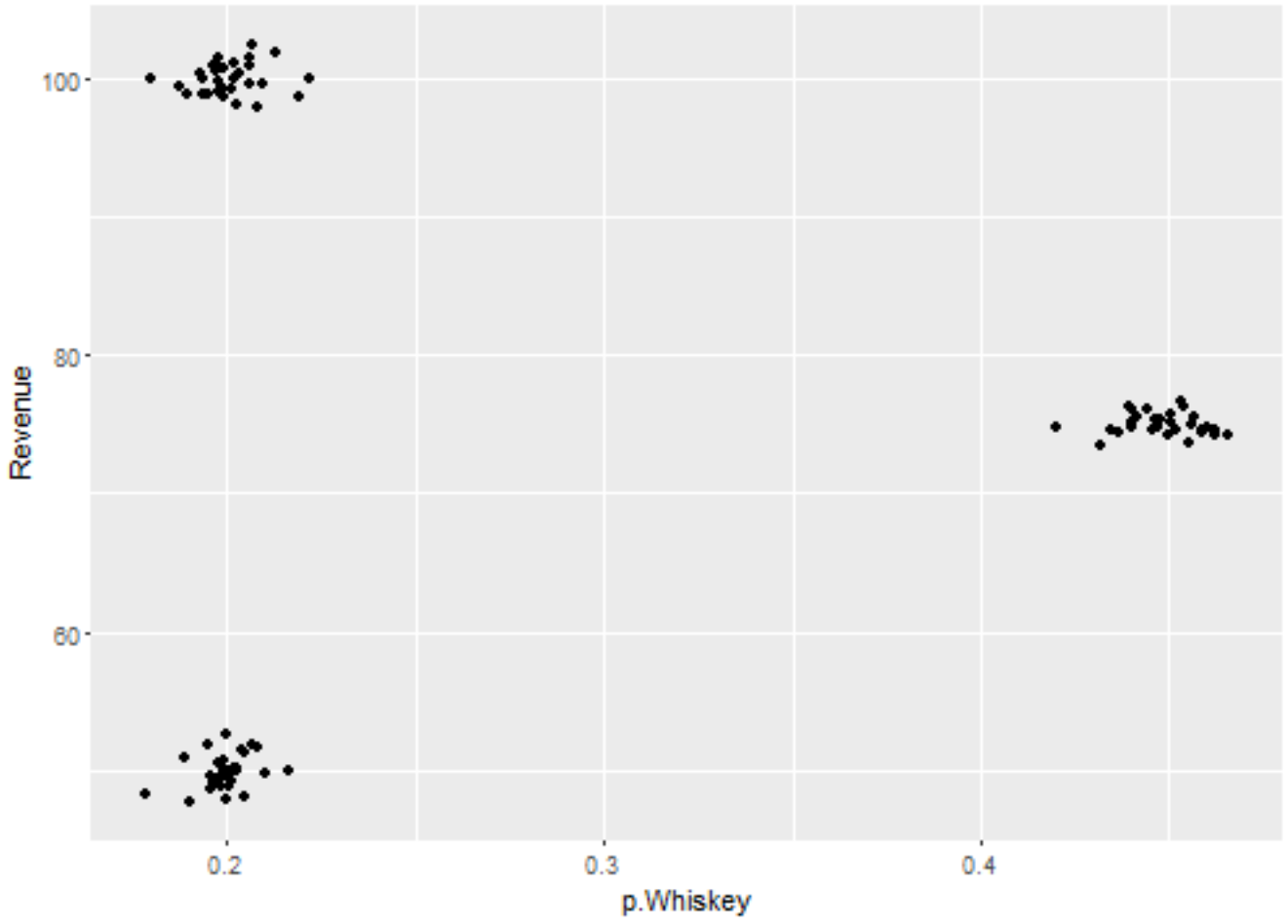
Unsupervised Learning

- Until now, there used to be a target variable.

Age	Income	Default
20	3000	1
30	4000	1
40	5000	0
50	6000	0
60	7000	0

- Sometimes, we don’t want to predict a target variable.

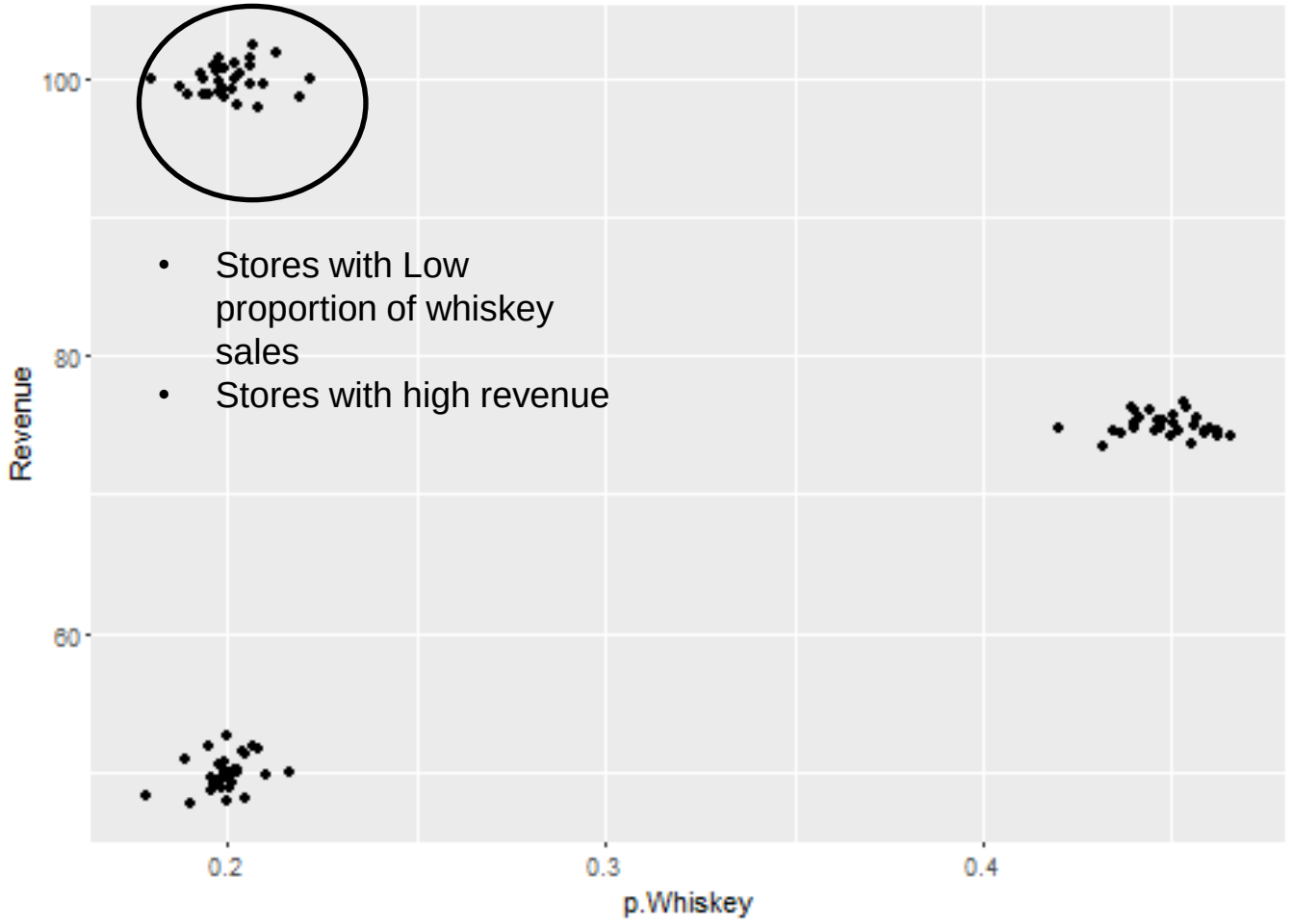
Unsupervised Learning



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

Can you discover any groups/clusters in this data?

Unsupervised Learning

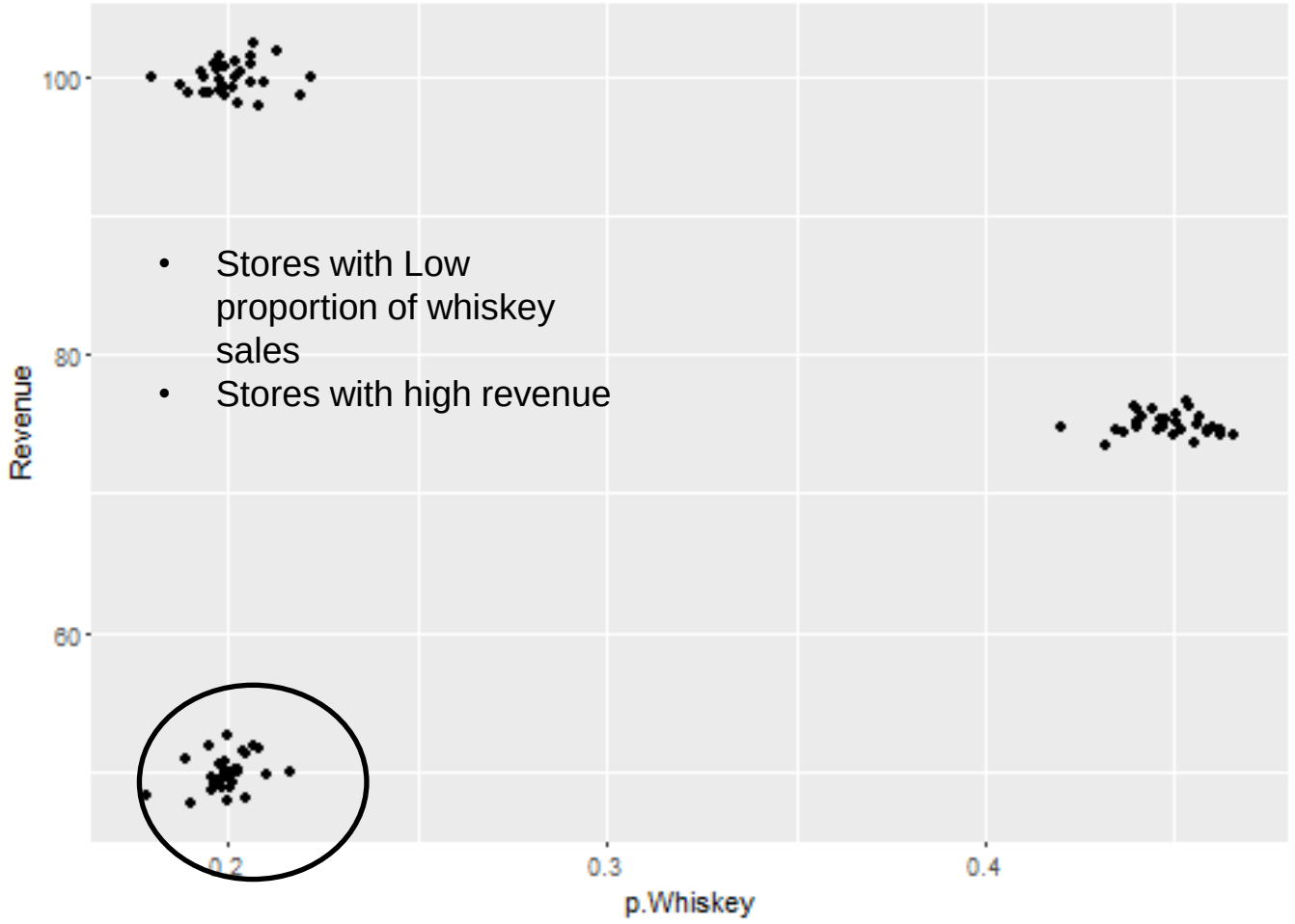


- Stores with Low proportion of whiskey sales
- Stores with high revenue

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

Can you discover any groups/clusters in this data?

Unsupervised Learning

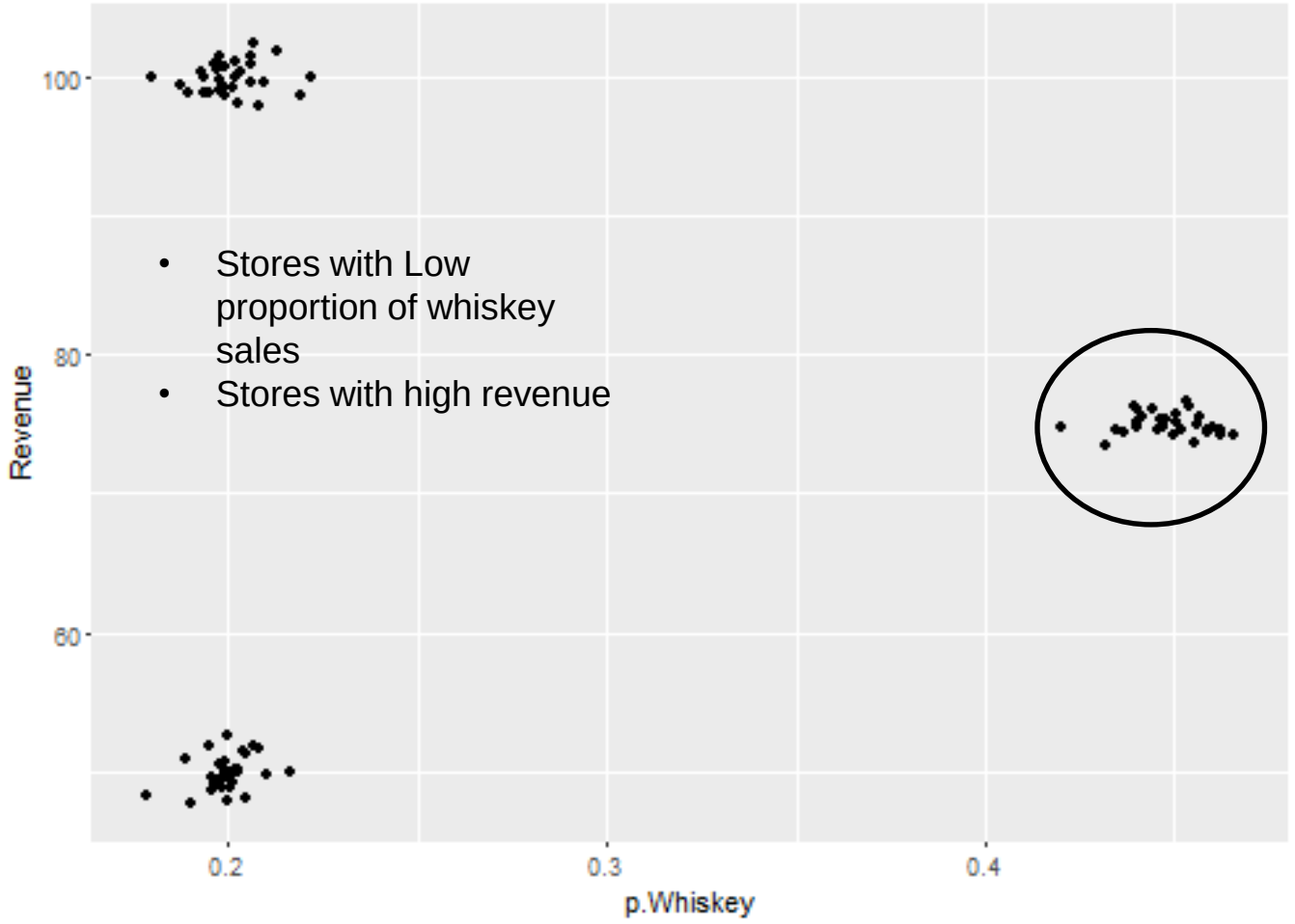


- Stores with Low proportion of whiskey sales
- Stores with high revenue

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

Can you discover any groups/clusters in this data?

Unsupervised Learning

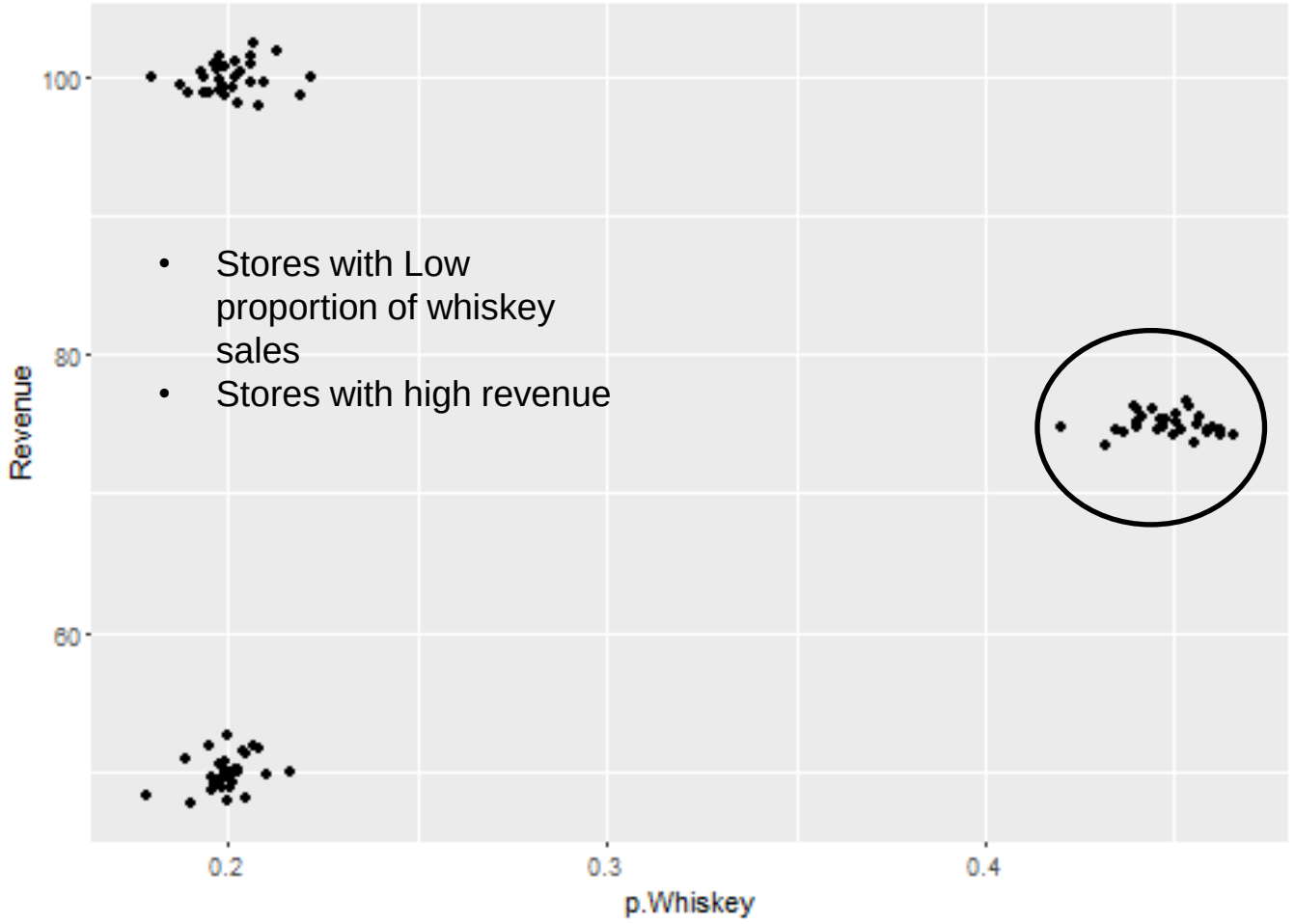


- Stores with Low proportion of whiskey sales
- Stores with high revenue

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

Can you discover any groups/clusters in this data?

Unsupervised Learning

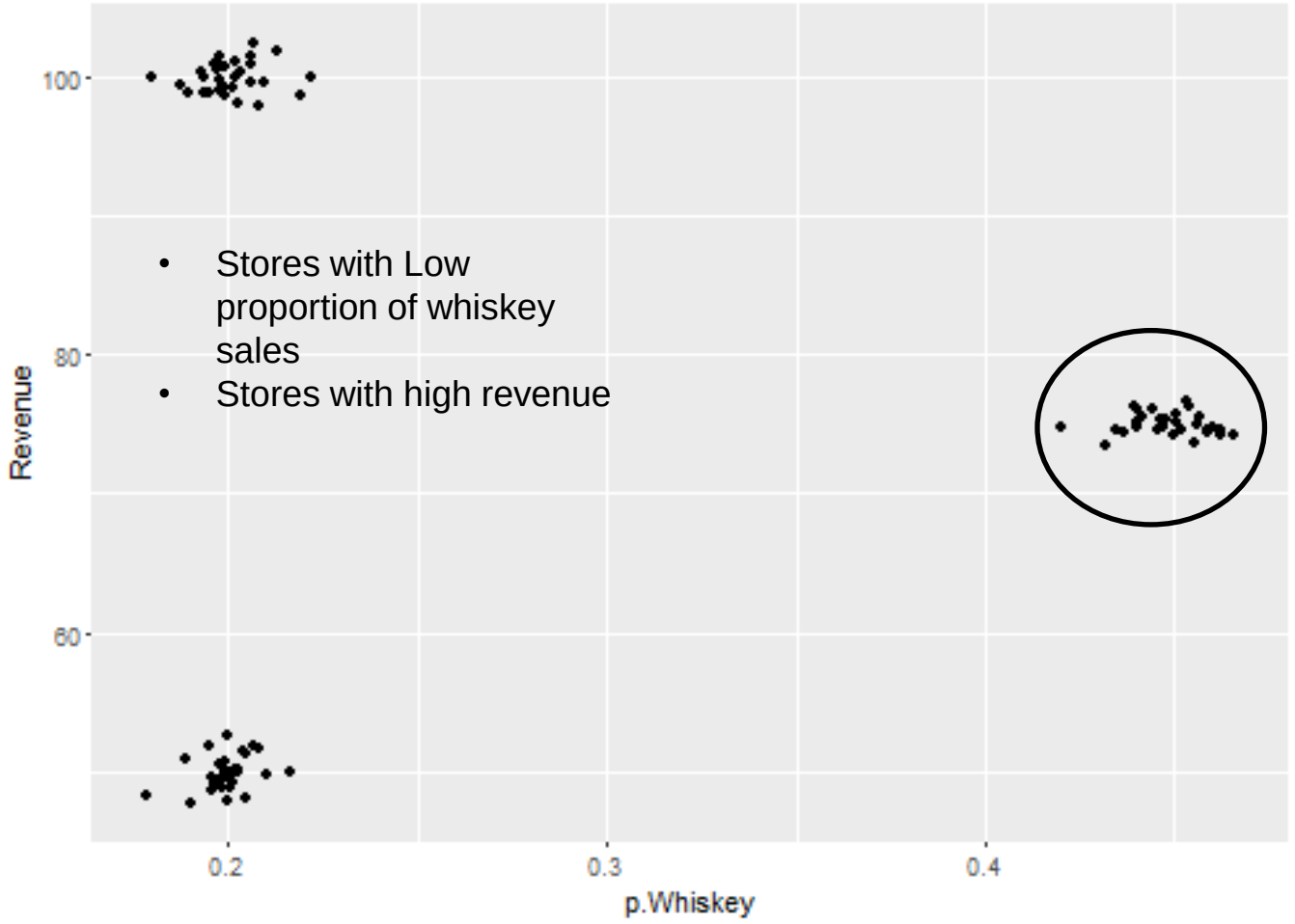


- Stores with Low proportion of whiskey sales
- Stores with high revenue

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

Is this task like the supervised learning we've seen before?

Unsupervised Learning



- Stores with Low proportion of whiskey sales
- Stores with high revenue

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

Is this task like the supervised learning we've seen before?

Is there any target variable we are trying to predict?

Unsupervised Learning: Class Exercise

- Suppose from a web analytics platform we have data on:
  - Number of hits
  - Duration of stays on a web page
  - Number of pages visited
  - Amount of money spent
- Discuss what type of segments you can discover?

Unsupervised Learning: Class Exercise

- Suppose from a web analytics platform we have data on:
  - Number of hits
  - Duration of stays on a web page
  - Number of pages visited
  - Amount of money spent
- Many categories can be potentially discovered
  - Group of people with a high hit rate, a small amount of money spent
  - Group of people with short duration stay, a large amount of money spent
  - And many more

Unsupervised Learning: Class Exercise

CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASE
C18647	1163.222817	0.909091	0.0	0.0	0.0
C16915	1778.035758	1.000000	449.0	449.0	0.0
C17090	837.631184	1.000000	0.0	0.0	0.0
C16927	1038.465061	1.000000	264.8	0.0	264.8
C15601	2269.073907	1.000000	284.9	284.9	0.0

Look at this data. Are there any business-relevant segments you can discover from this dataset?

Hero

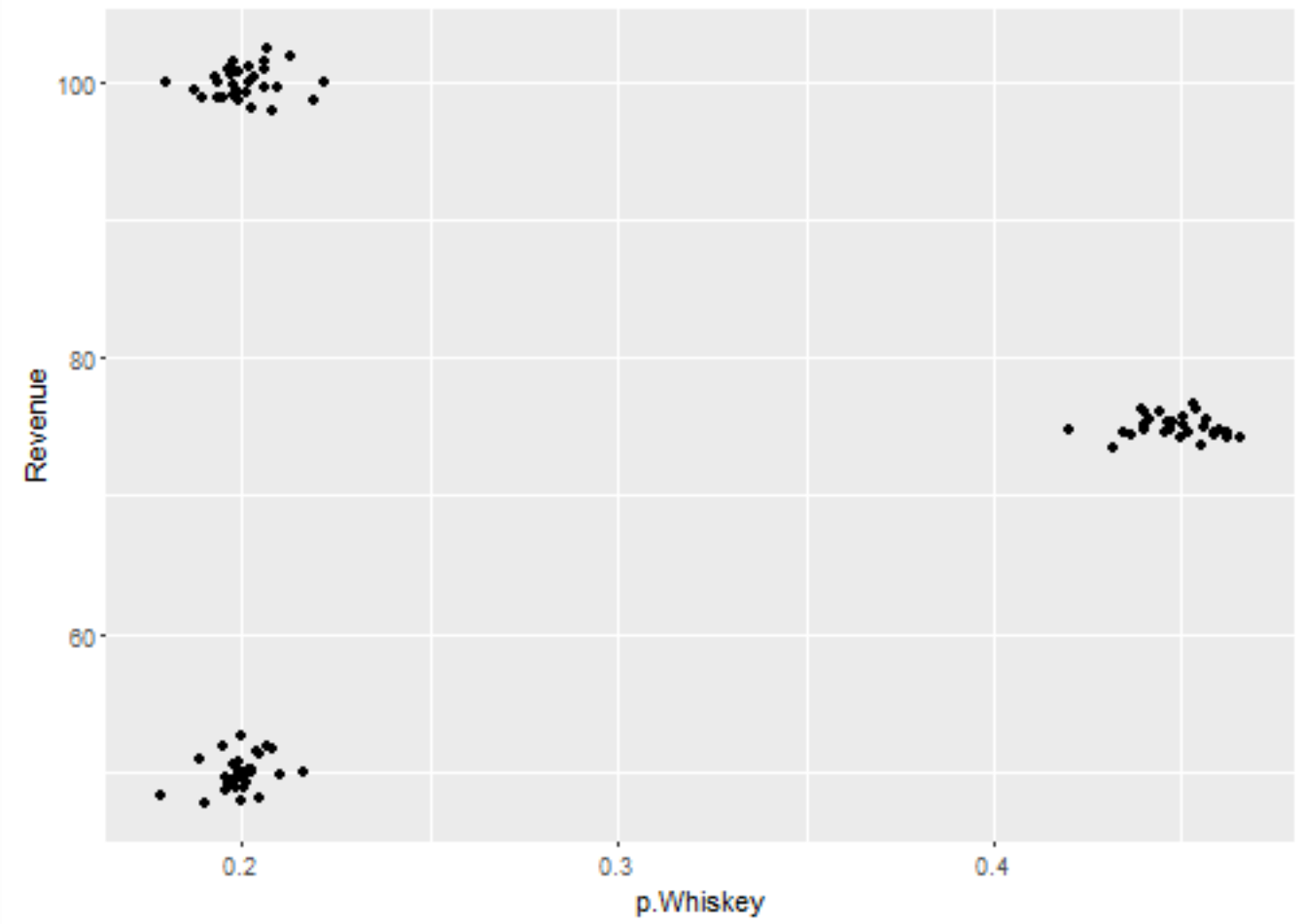
Unsupervised Learning: Class Exercise

CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASE
C18647	1163.222817	0.909091	0.0	0.0	0.0
C16915	1778.035758	1.000000	449.0	449.0	0.0
C17090	837.631184	1.000000	0.0	0.0	0.0
C16927	1038.465061	1.000000	264.8	0.0	264.8
C15601	2269.073907	1.000000	284.9	284.9	0.0

Look at this data. Are there any business-relevant segments you can discover from this dataset?

- Group of customers with high balance but meager purchase amounts (Potential for upsell, offers)
- Group of customers with low balance but very high purchase amounts (Potentially risky customers)

Need for Clustering Algorithm



- How many groups?
- How did you find it?
- What if you need segments by four variables?

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

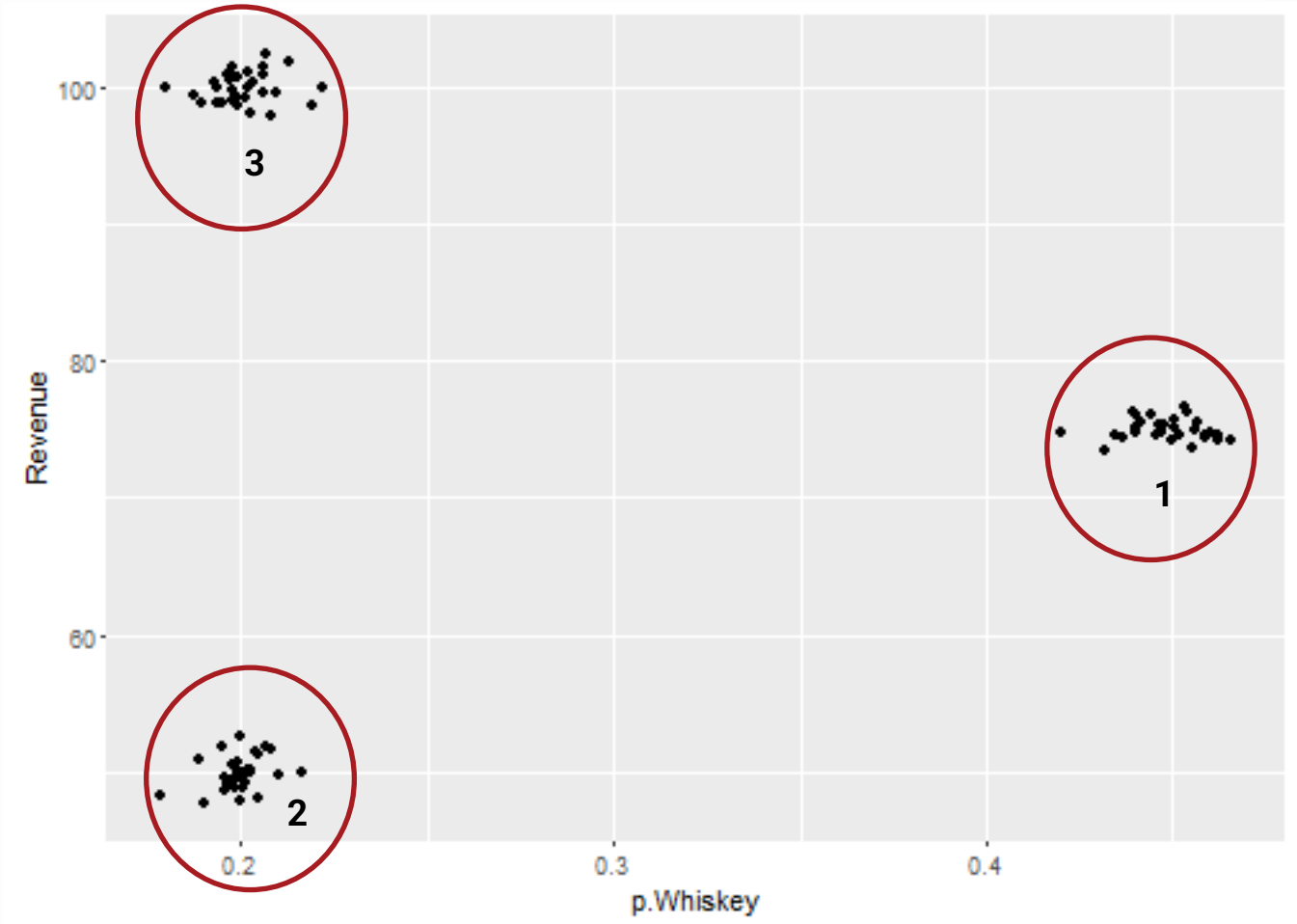
Need for Clustering Algorithm

- Cluster: How? Visualize?
- What if eight variables are used that we need to create clusters?
- We need an algorithm
- KMeans

Clustering: KMeans

- **KMeans** is an iterative algorithm.
- The “K” in **KMeans**, stands for the number of clusters to be found in data; it’s a user-supplied parameter.
- The output of the **KMeans** algorithm is the cluster label for each row of data.

Clustering: KMeans



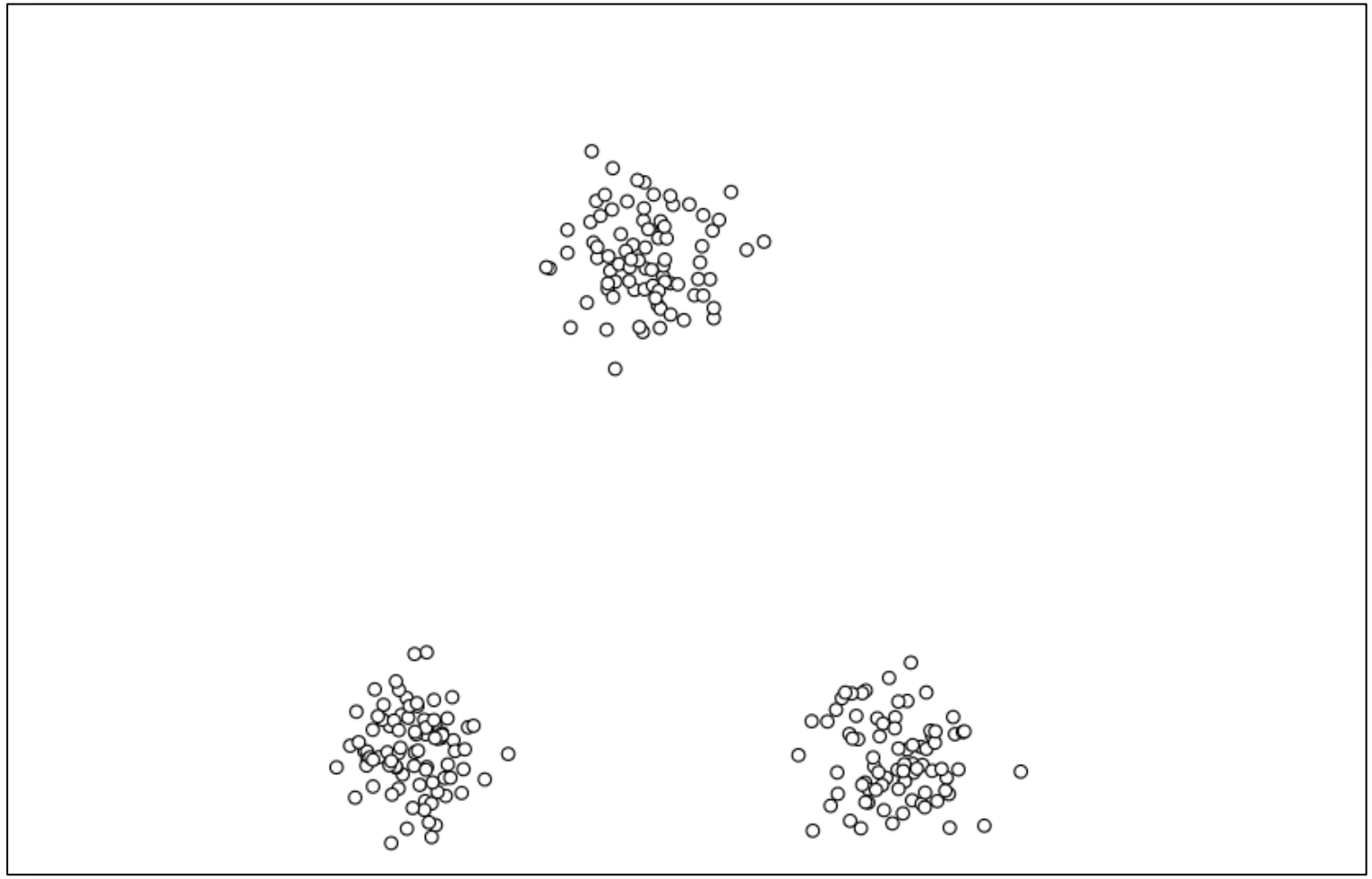
P.Whisky	Revenue	Label
0.40	80	1
0.20	60	2
0.35	40	2
0.22	90	3
0.45	75	1

Clustering: KMeans

Algorithm

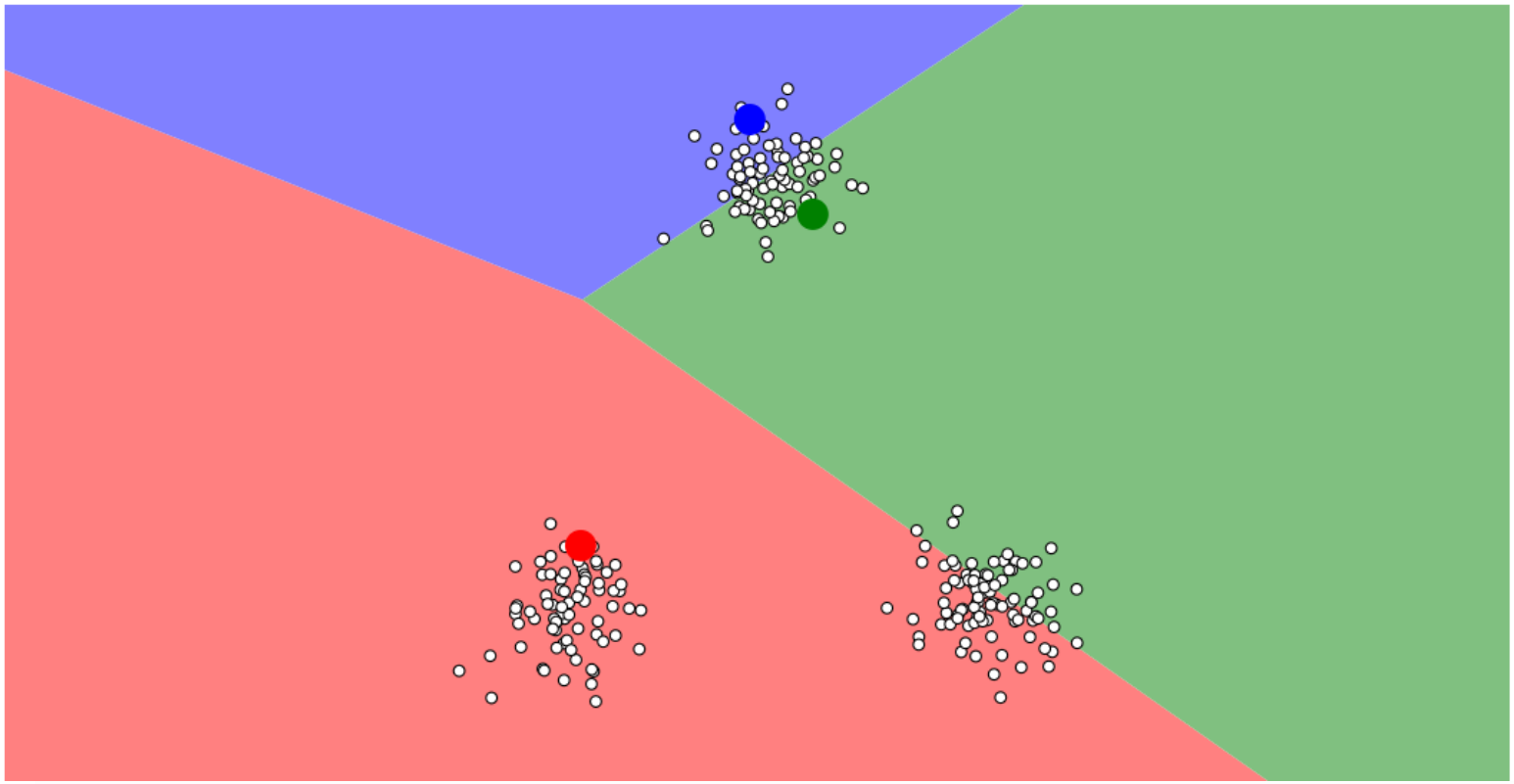
1. Initialize “K” cluster centers randomly
   a) Find the distance of each point from the “K” cluster centers.
   b) Based on the proximity of each point from the “K” cluster centers, we give cluster labels to each data point.
   c) After the clusters are formed, we re-compute the cluster centers.
   d) Repeat a), b), and c) till convergence (till no data point changes cluster membership in succeeding iterations).

Clustering: KMeans



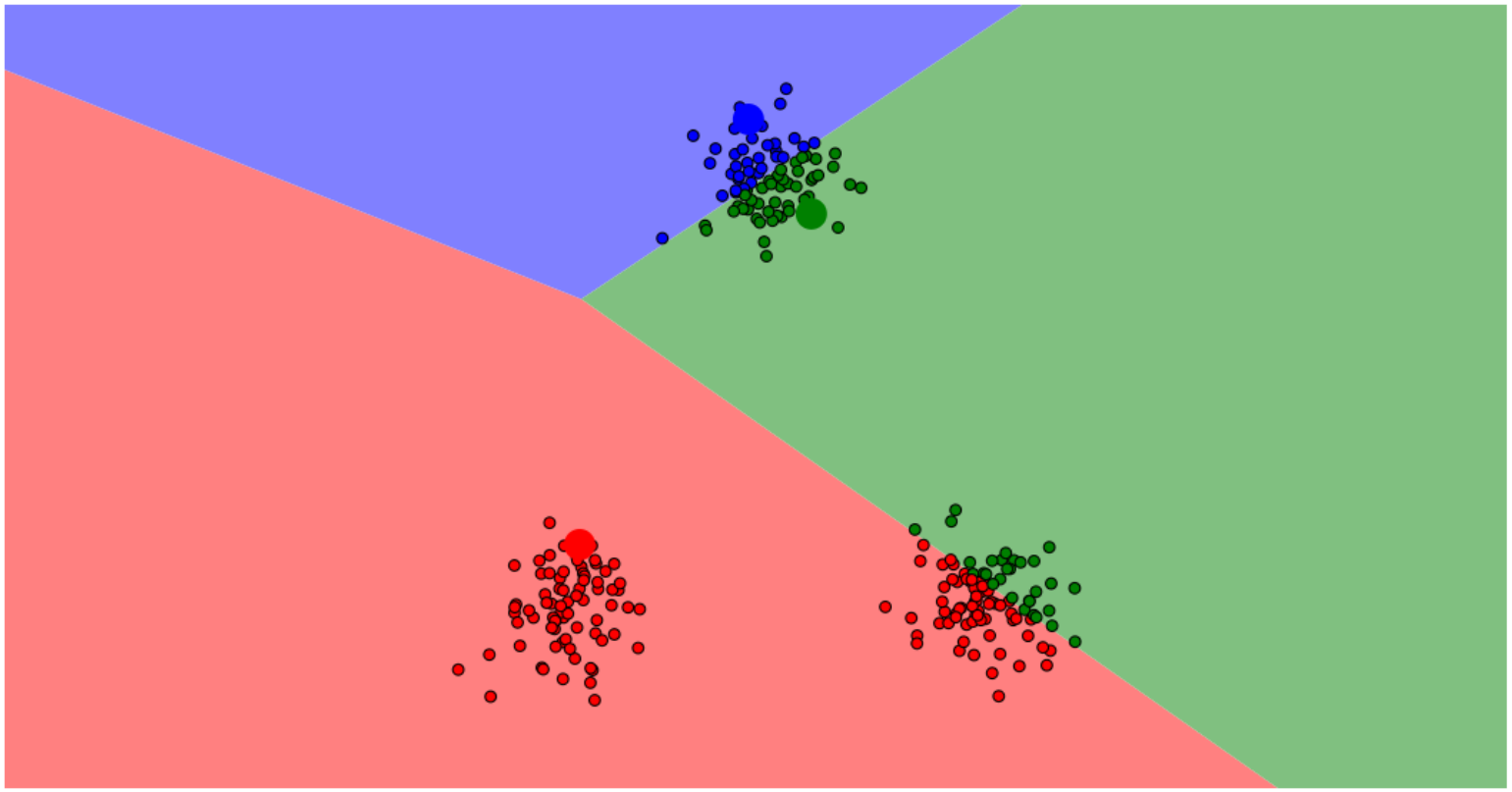
There are 3 data clusters, K = 3.

Clustering: KMeans



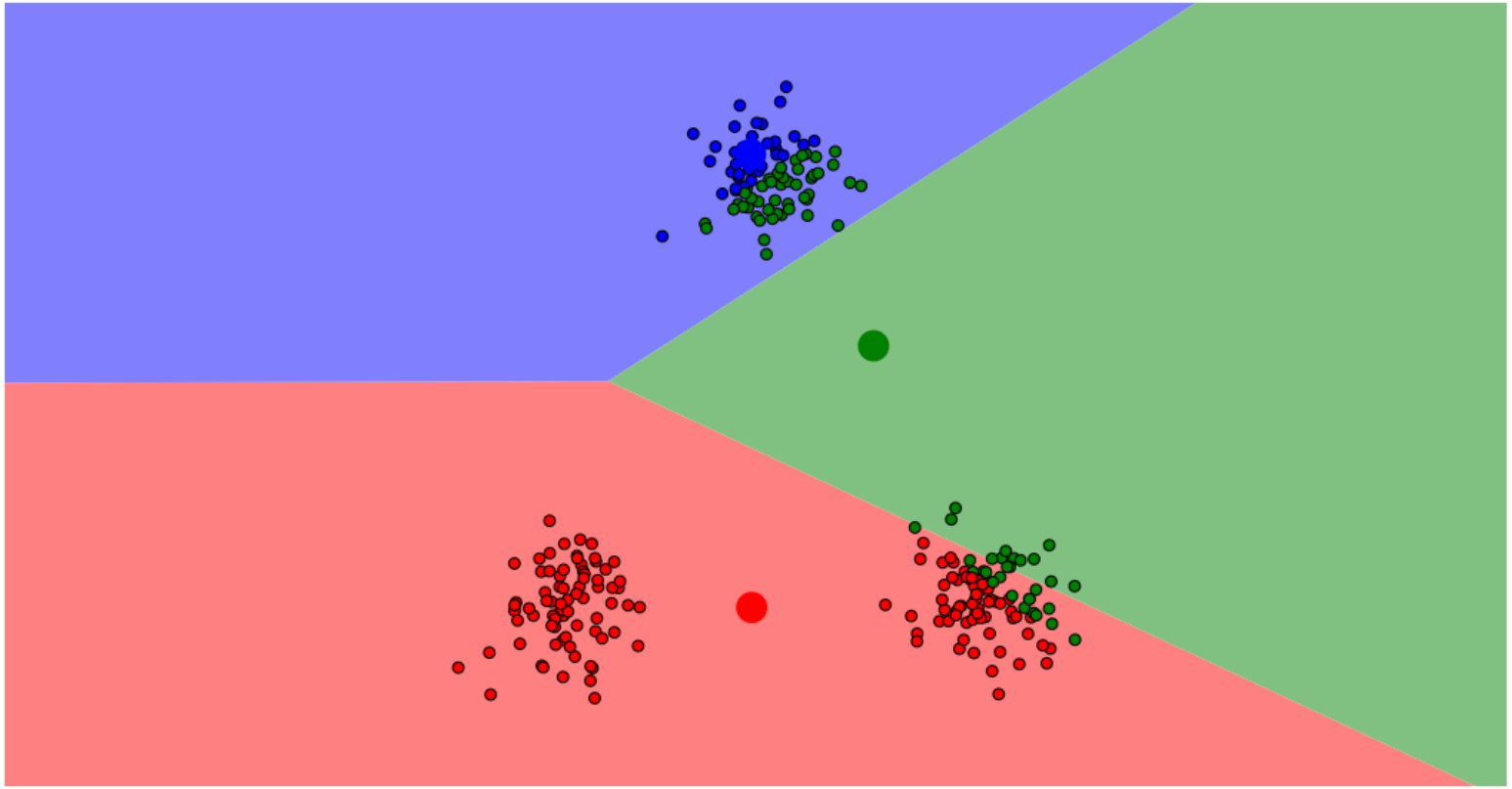
Randomly assign 3 points as **cluster centers**.

Clustering: KMeans



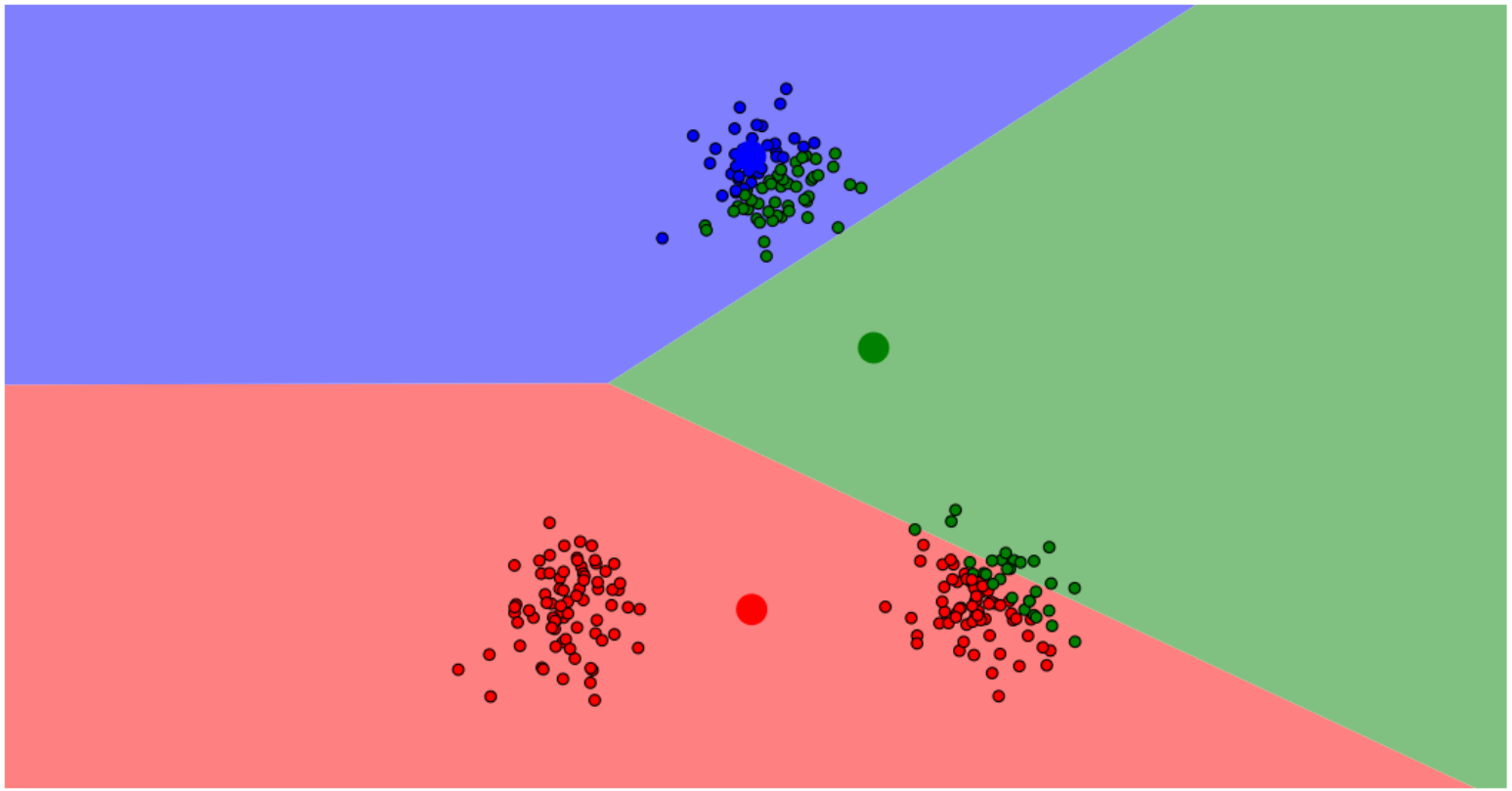
Compute the distances of each point from each of the cluster centers and assign cluster labels.

Clustering: KMeans



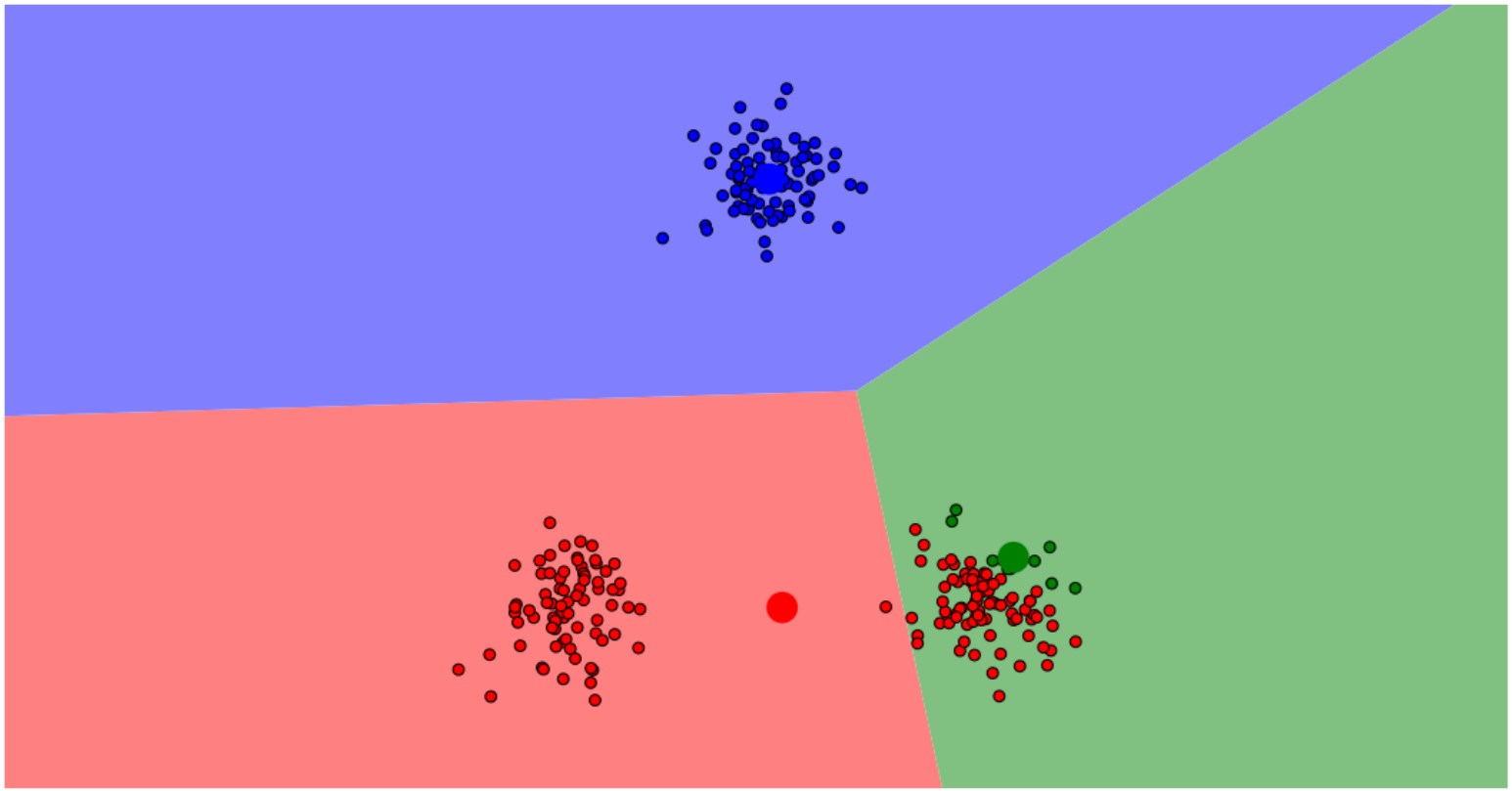
Re-compute cluster centers.

Clustering: KMeans



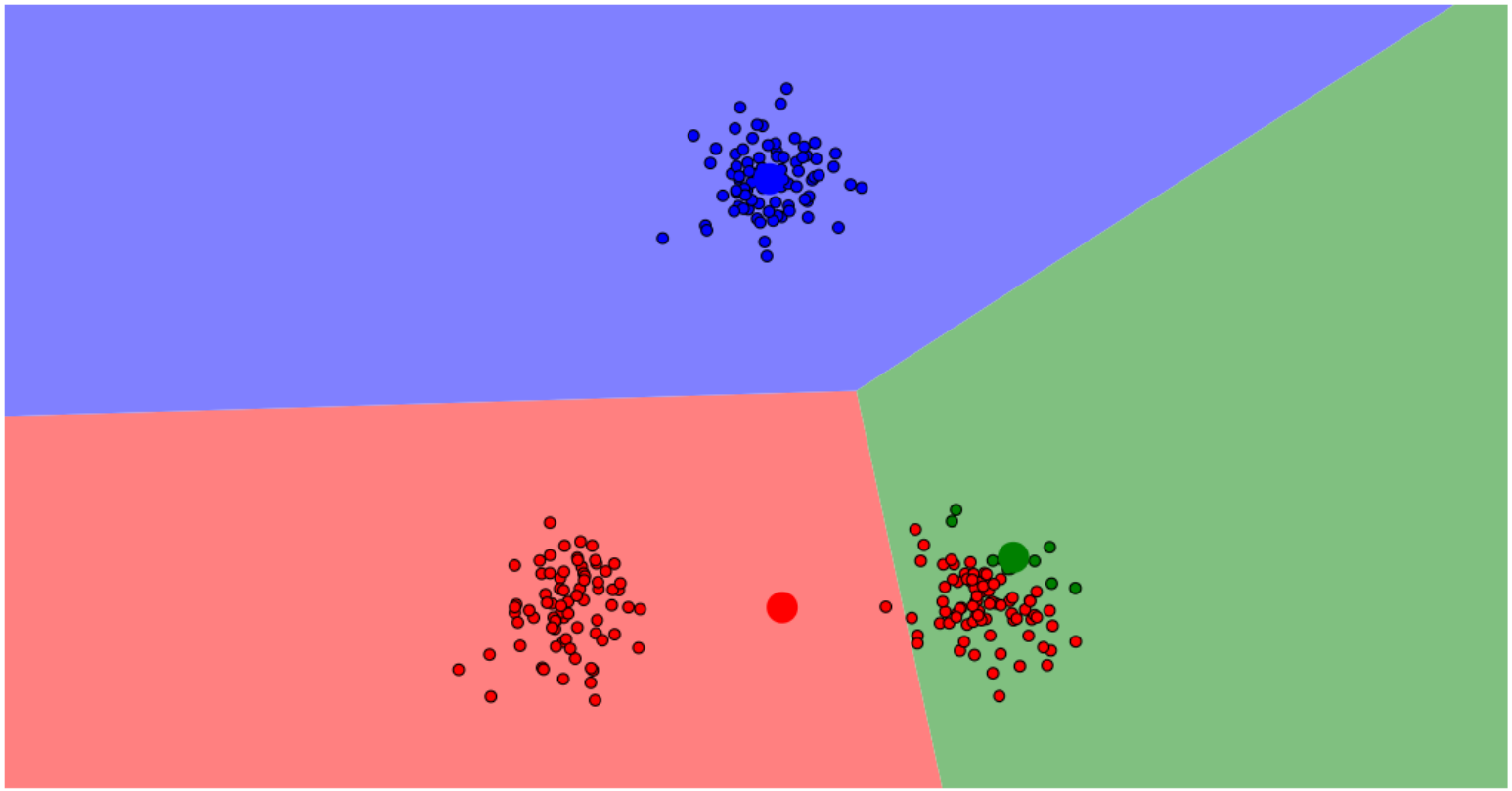
Compute distances of each point from each of the cluster centers and assign cluster labels

Clustering: KMeans



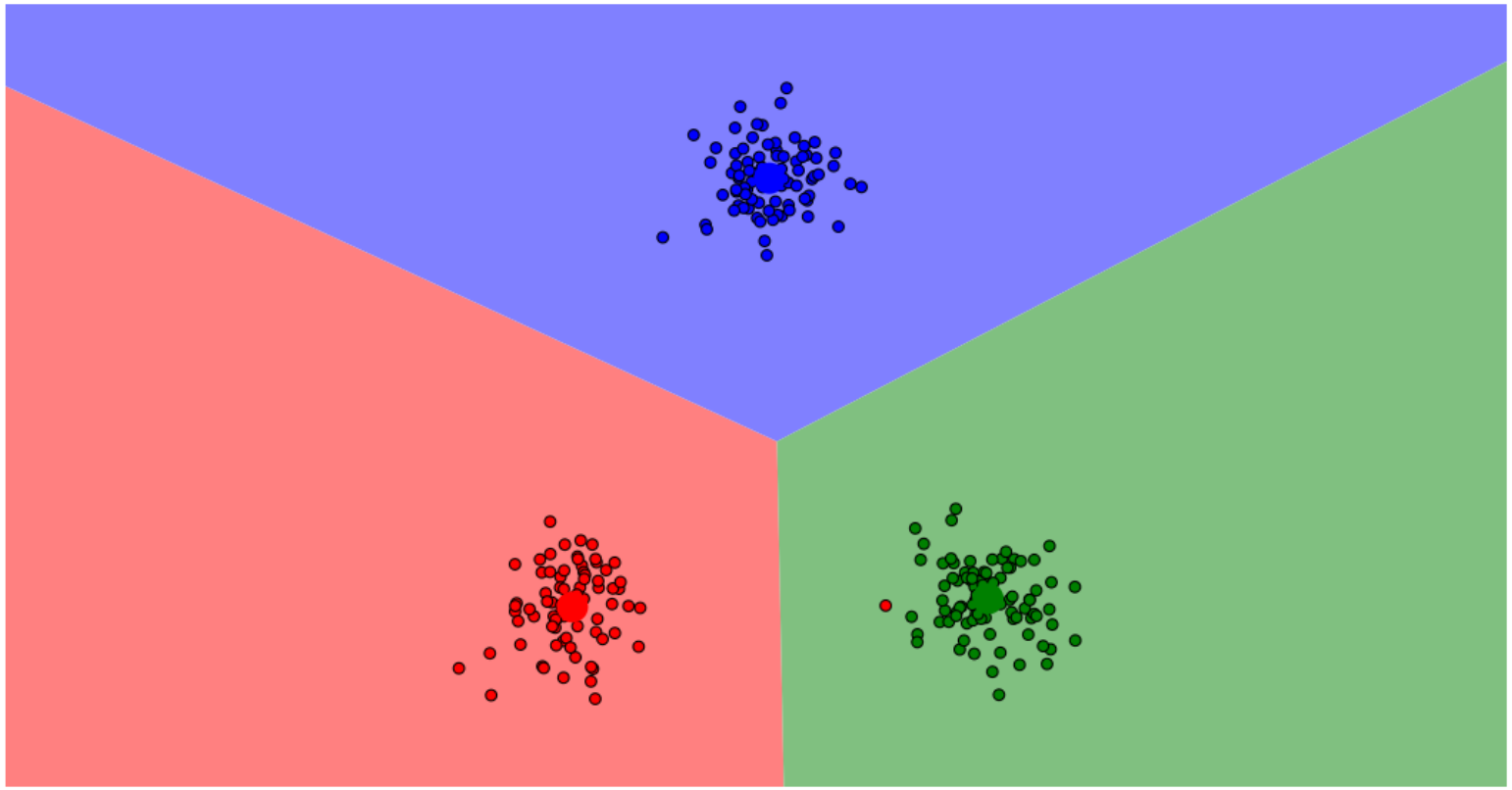
Recompute cluster centers.

Clustering: KMeans



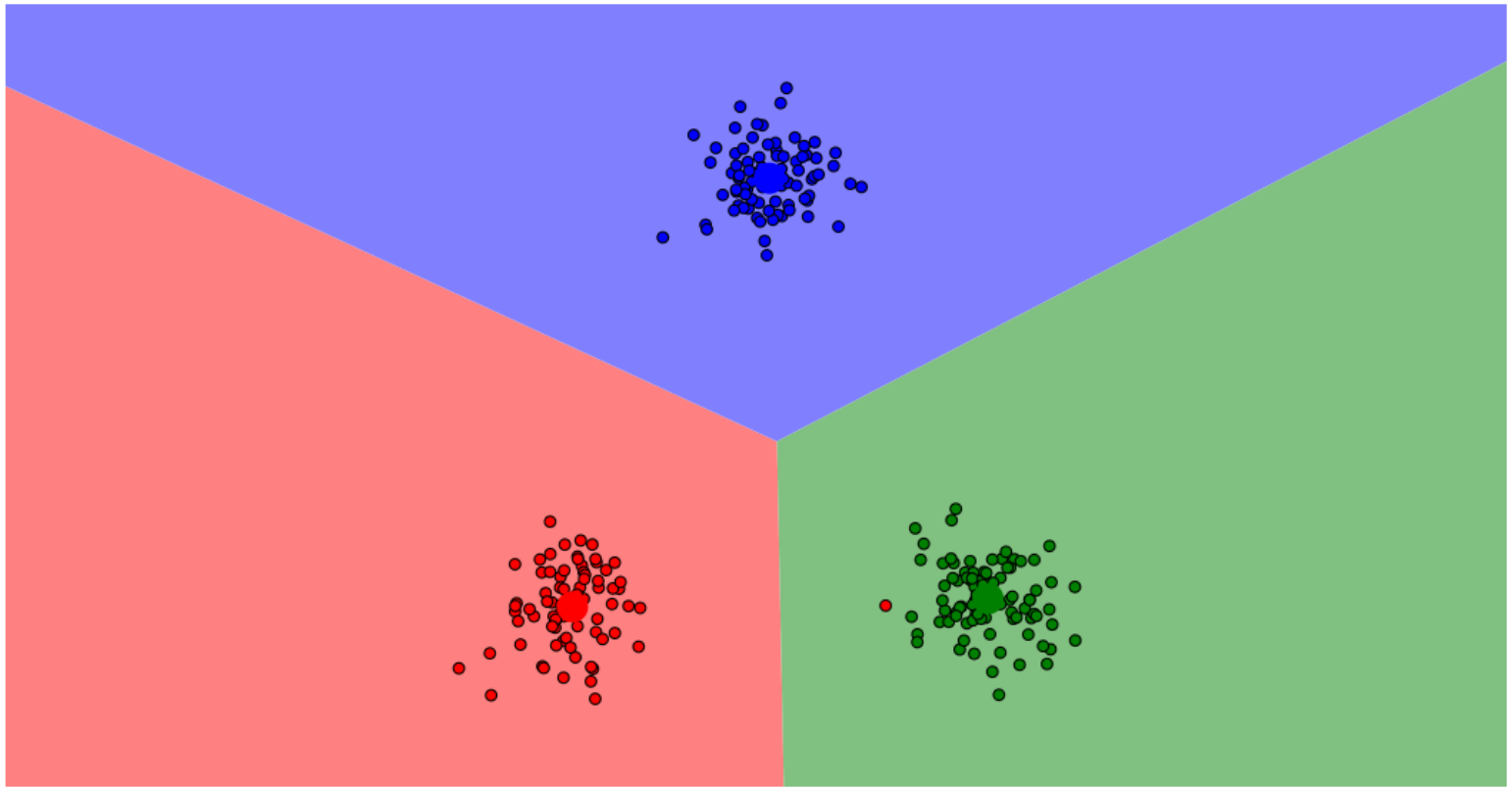
Compute the distances of each point from each of the cluster centers and assign cluster labels.

Clustering: KMeans



Recompute cluster centers.

Clustering: KMeans



Compute the distances of each point from each of the cluster centers and assign cluster labels.

Clustering: KMeans

Demonstrate **Numerical Example Clustering.xlsx**

Please go through the excel sheet before the class.

Clustering Class Exercise

Age	Weight
30	67
35	70
25	64
23	62

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

Find out which clusters each row will belong to. **Compute** the distance of each point from the centers and then decide.

Clustering Class Exercise

Age	Weight
30	67
35	70
25	64
23	62

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

Find out which clusters each row will belong to. **Compute** the distance of each point from the centers and then decide.

Solution: See **Numerical Example Clustering.xlsx**, sheet named **Class Exercise.**

Clustering: KMeans

Once we intuitively understand how **Kmeans** work, there are some peculiarities about **Kmeans** that we need to keep in mind:

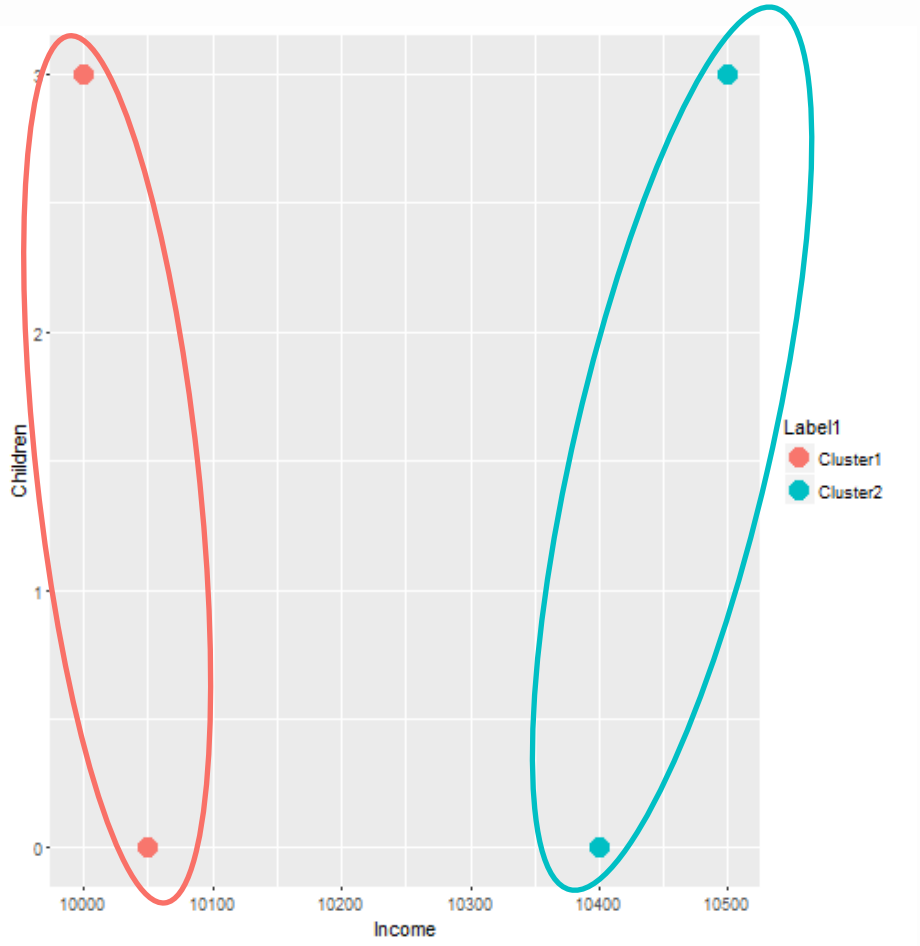
- **Data Level:**
  - Only numeric data can be fed to a **Kmeans** algorithm.
  - Data should be scaled
- **Algorithm Level:**
  - How to find out what could be a good value of “K”?

Clustering: Kmeans

Revenue	Size (Sq ft)	Footfall	City
4000	1000	80	Blr
3000	1200	90	Blr
8000	1400	100	Chennai
9000	900	200	Blr
2000	1234	324	Chennai

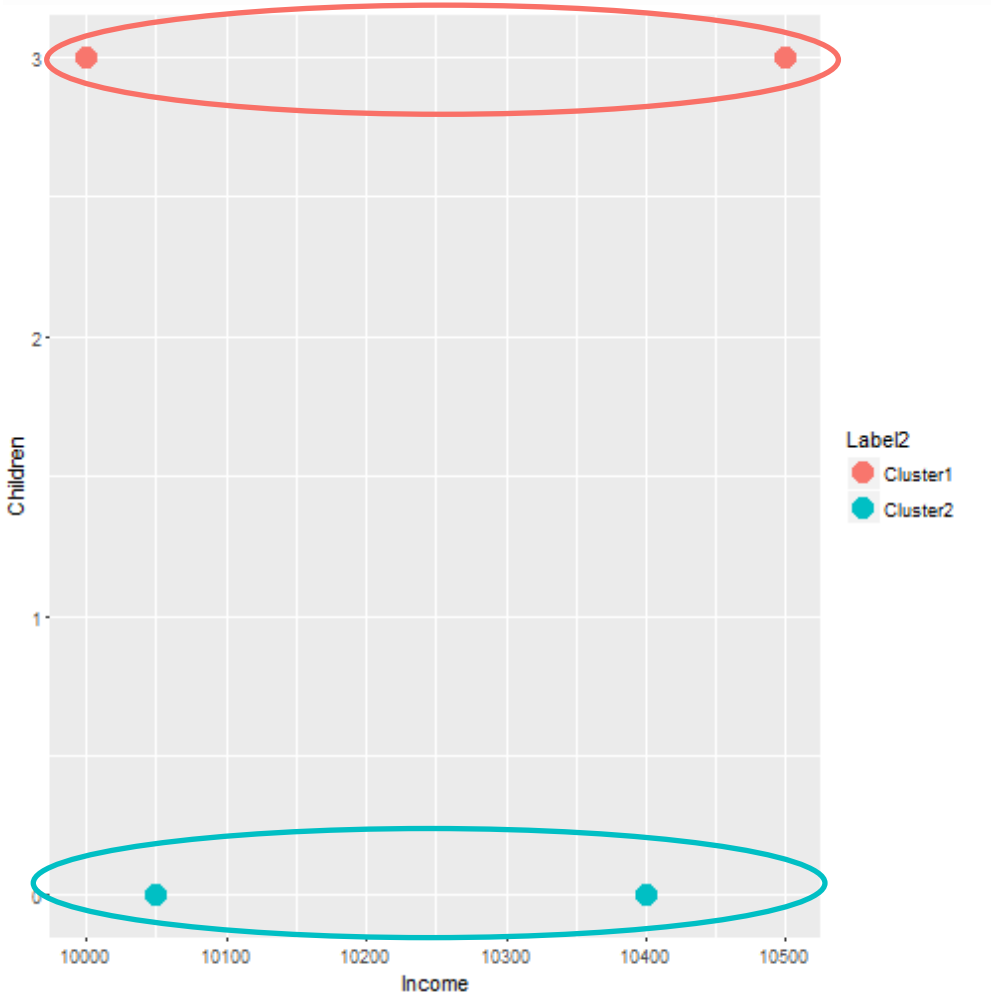
Revenue	Size (Sq ft)	Footfall	City_Blr	City_Ch
4000	1000	80	1	0
3000	1200	90	1	0
8000	1400	100	0	1
9000	900	200	1	0
2000	1234	324	0	1

Clustering: Kmeans, Data Should be Scaled



Income	Children
10050	0
10050	2
10400	0
10500	2

Clustering: Kmeans, Data Should be Scaled



Income	Children
10050	0
10050	2
10400	0
10500	2

Clustering: Kmeans, Data Should be Scaled

Income	Children
10050	0
10050	2
10400	0
10500	2

$$Z_i = (x_i - \mu)/\sigma$$

$$\mu_{Income} = 10250$$
$$\sigma_{Income} = 203.10$$

$$\mu_{children} = 1$$
$$\sigma_{children} = 1$$

Income	Children
-0.98	-1
-0.98	1
0.73	-1
1.73	1

Similar Scale

Z scores are not the only way to standardize data; one can use min max scaler link and absolute scaler link.

Clustering Class Exercise

Look at the data and discuss what kind of transformations you will need to do for the variables present here.

Age	Graduated	Profession	Work_Exp...	Spending_...	Family_Size
36	Yes	Engineer	0.0	Low	1.0
37	Yes	Healthcare	8.0	Average	4.0
69	No		0.0	Low	1.0
59	No	Executive	11.0	High	2.0
19	No	Marketing		Low	4.0
47	Yes	Doctor	0.0	High	5.0
61	Yes	Doctor	5.0	Low	3.0
47	Yes	Artist	1.0	Average	3.0
50	Yes	Artist	2.0	Average	4.0
19	No	Healthcare	0.0	Low	4.0
22	No	Healthcare	0.0	Low	3.0
22	No	Healthcare	0.0	Low	6.0
50	Yes	Artist	1.0	Average	5.0
27	No	Healthcare	8.0	Low	3.0
18	No	Doctor	0.0	Low	3.0
61	Yes	Artist	0.0	Low	1.0

Clustering Class Exercise

Look at the data and discuss what kind of transformations you will need to do for the variables present here.

- **Age:** Standardize
- **Graduated:** Either drop or create a dummy
- **Profession:** Either drop 9 or create a dummy
- **Work Experience:** Standardize
- **Spending:** Either drop or create a dummy
- **Family Size:** Standardize

Age	Graduated	Profession	Work_Exp...	Spending_...	Family_Size
36	Yes	Engineer	0.0	Low	1.0
37	Yes	Healthcare	8.0	Average	4.0
69	No		0.0	Low	1.0
59	No	Executive	11.0	High	2.0
19	No	Marketing		Low	4.0
47	Yes	Doctor	0.0	High	5.0
61	Yes	Doctor	5.0	Low	3.0
47	Yes	Artist	1.0	Average	3.0
50	Yes	Artist	2.0	Average	4.0
19	No	Healthcare	0.0	Low	4.0
22	No	Healthcare	0.0	Low	3.0
22	No	Healthcare	0.0	Low	6.0
50	Yes	Artist	1.0	Average	5.0
27	No	Healthcare	8.0	Low	3.0
18	No	Doctor	0.0	Low	3.0
61	Yes	Artist	0.0	Low	1.0

Clustering Kmeans: Python Demo

Github link to demo [link]

Thank You!