MACHINE LEARNING

Algorithms

Class

Recommendation Engines

Topic

Introduction to Clustering: K-means

Agenda

1. **Unsupervised machine learning algorithms**
2. **Clustering** – a class of machine learning algorithm
3. 2 clustering algorithms – **K-means** and **Agglomerative Clustering**
4. Algorithms will be explained in context of **business data** and how to **segment data using these algorithms**
5. Use clustering in an **unsupervised image classification task**

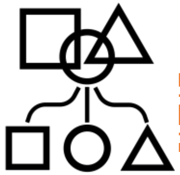
Unsupervised Learning

Unsupervised learning

Versus

Supervised learning

A variable that we predict



Typical classification problem → Predict the default status of a customer

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

The task can be accomplished using many Supervised learning classifiers like logistic regression, trees, random forests etc.

Unsupervised Learning

Unsupervised learning

Versus

Supervised learning

Finding interesting patterns in the data to make predictions or take business actions based on that, without predicting any variables

A variable that we predict

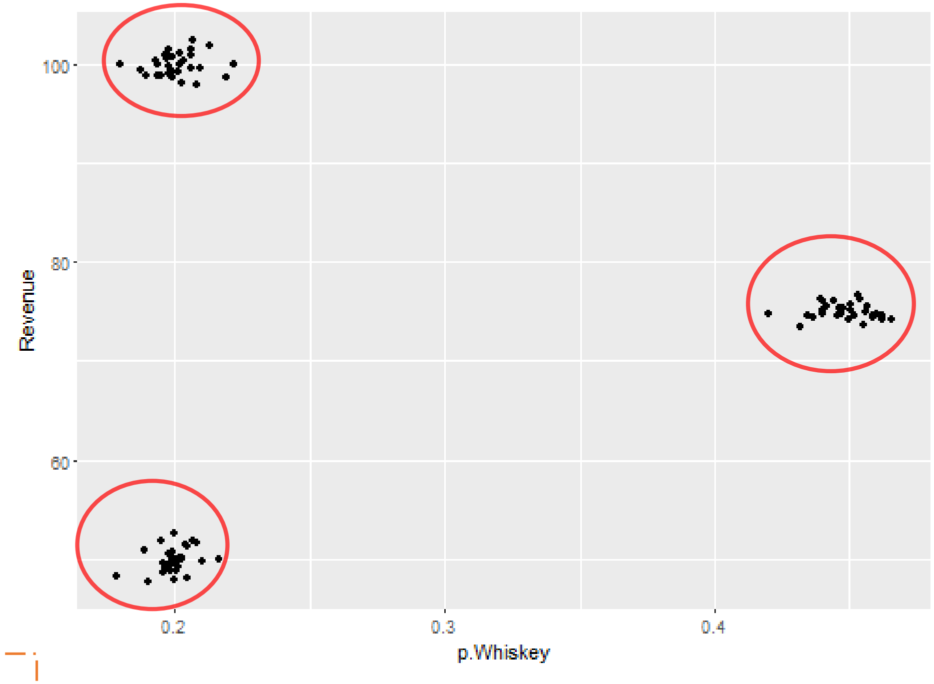
Store_id	P.Whisky	Revenue
1	0.40	80
2	0.20	60
3	0.35	40
4	0.22	90
5	0.45	75

3 categories of stores:

a. Stores with low revenue and low percentage of whiskey sales

b. Stores with high percentage of whiskey sales and the revenue is neither too low nor too high

c. Stores with high revenue and low percentage of whiskey sales



Unsupervised Learning

Unsupervised learning | Versus | Supervised learning

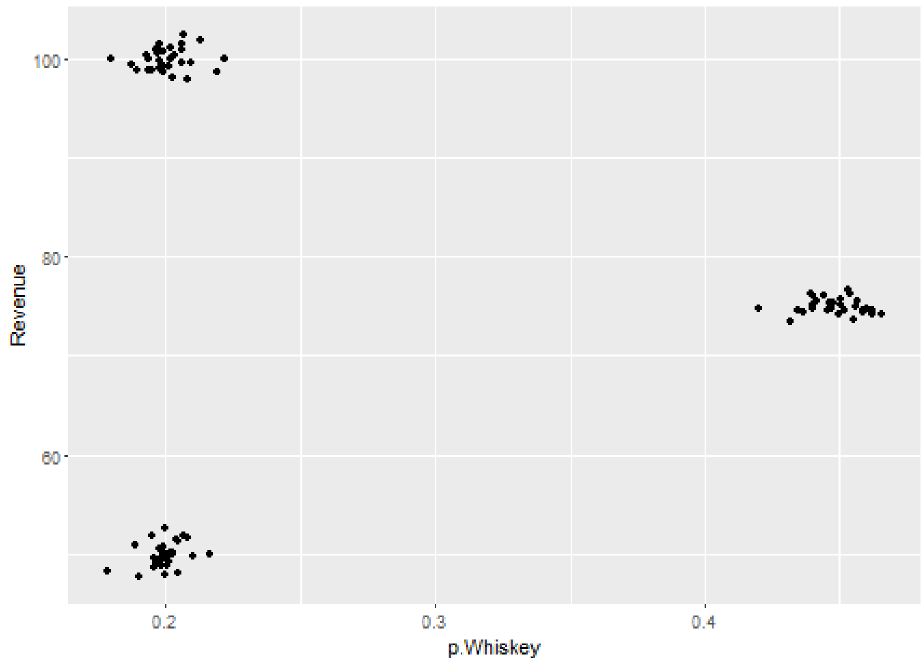
Finding interesting patterns in the data to make predictions or take business actions based on that, without predicting any variables

A variable that we predict

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

Identified 3 interesting categories of stores from the data but no target variable was predicted

No classification task or regression task was done



Example of Unsupervised learning → Clustering

Clustering

Clustering → Find out interesting groups in a numeric data

Other functions of clustering

Web analytics platform

Capture data of people activity on the webpage

Duration of stays on the page

The number of pages visited

Number of hits

Many categories can be potentially discovered

High hit rate

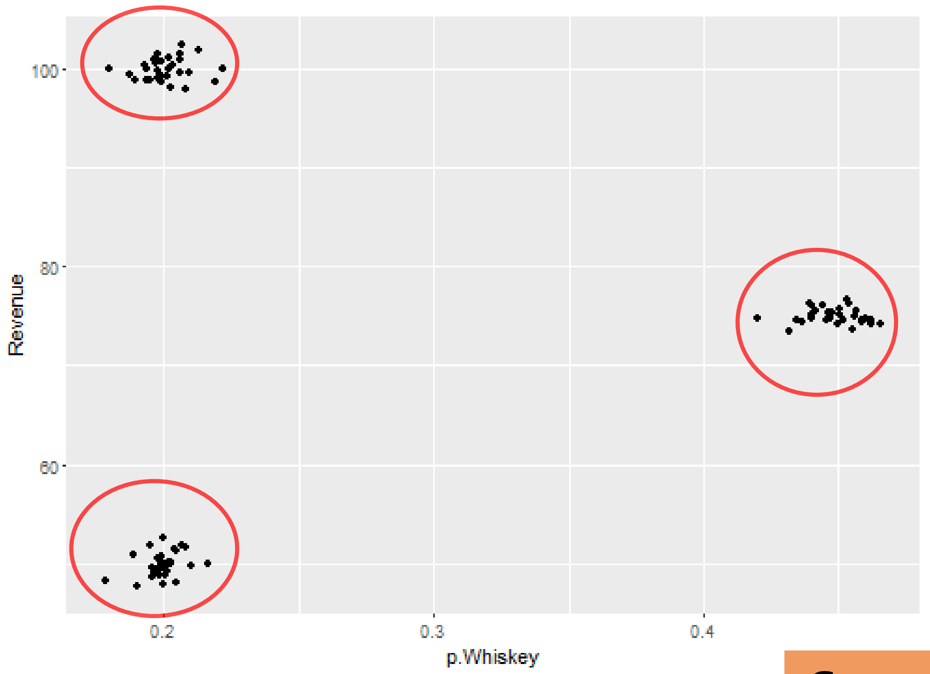
Short duration stay

Amount of money spent

Small

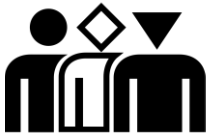
Large

Clustering



Visual representation

3 groups

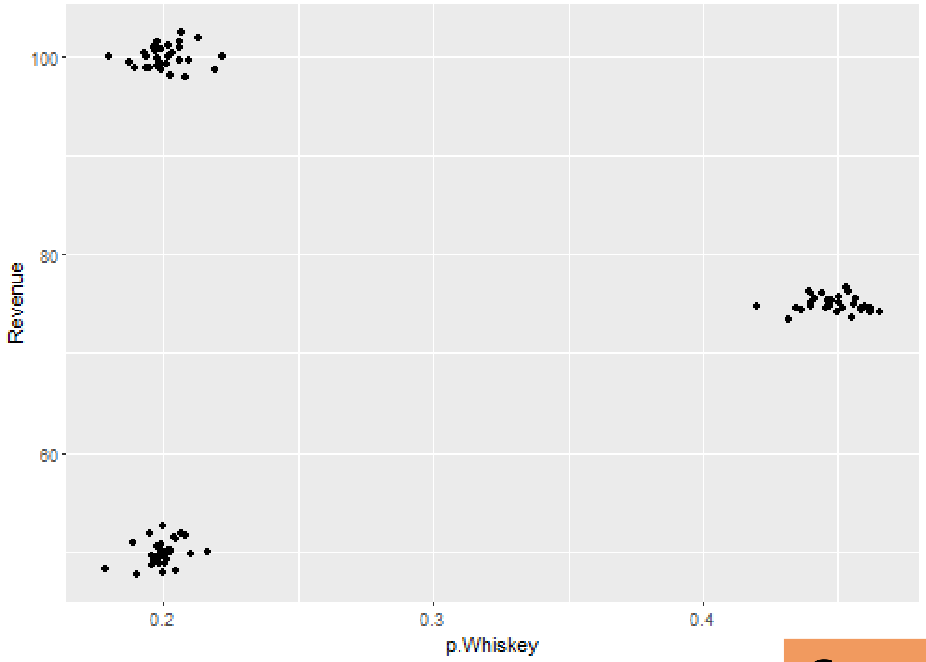


2 variables

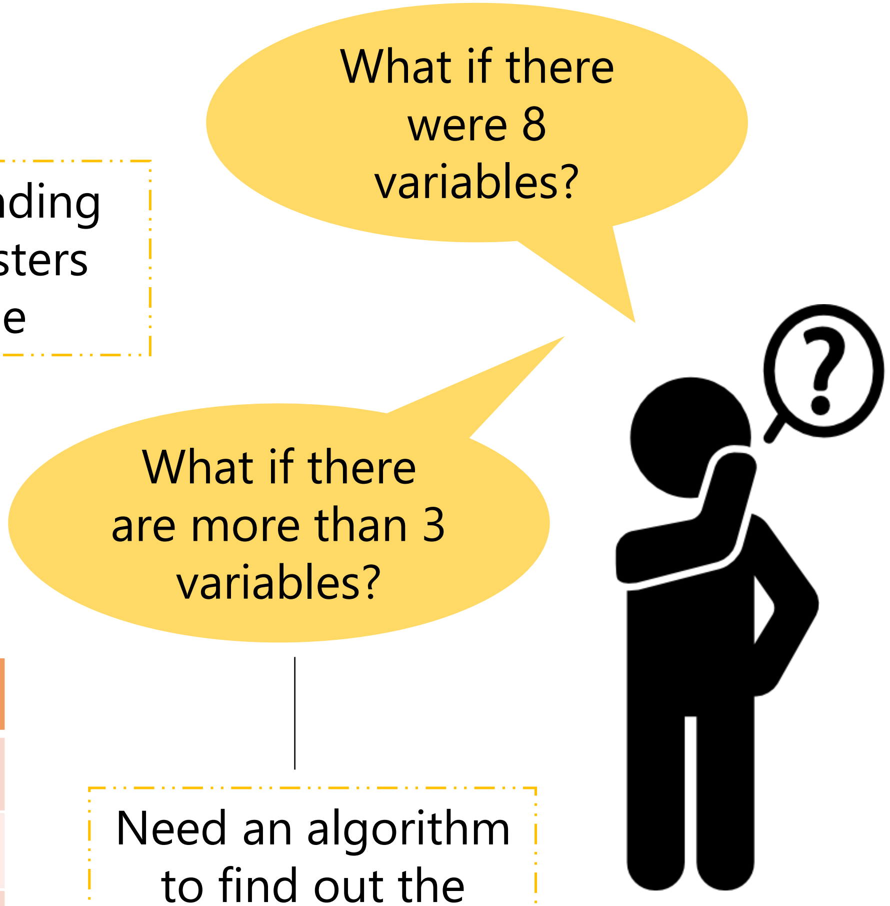
Retail store data

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

Clustering



Visualizing data and finding out the number of clusters would be impossible



Retail store data

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 an algorithm to find out the clusters in our data

Clustering – K-means

What if there were 8 variables?

What if there are more than 3 variables?

Need an algorithm to find out the clusters in our data

Popular algorithms

K-means

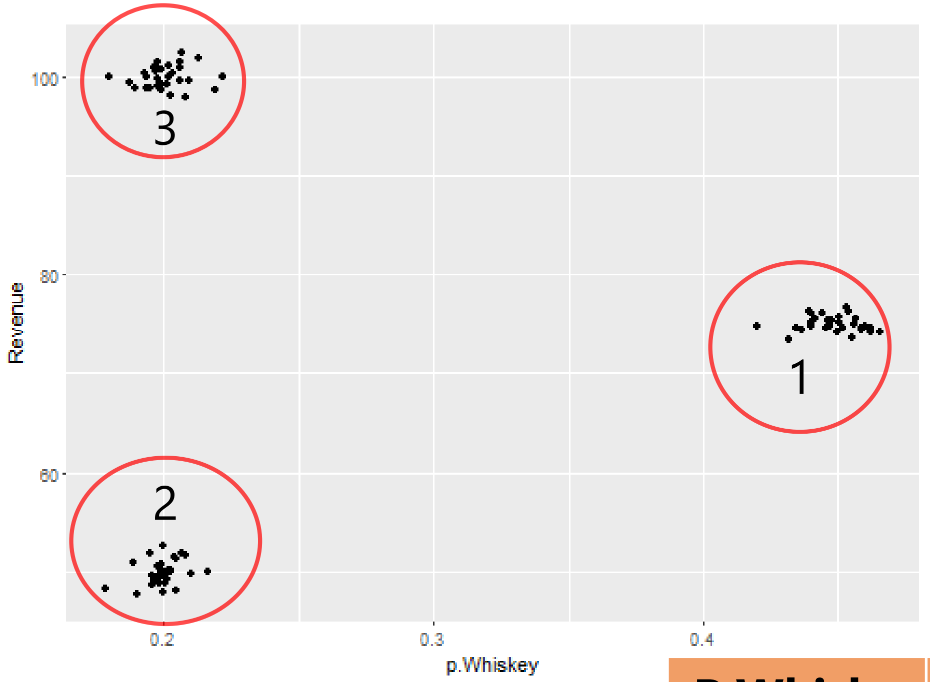
Output of K-means algorithm is the cluster label for each row of the data

User specified parameter

K in K-means stands for the number of clusters to be found out in data

K-means - an iterative algorithm

Clustering – K-means



When any clustering algorithm such as K-means is used

End up finding out which data point belongs to which group

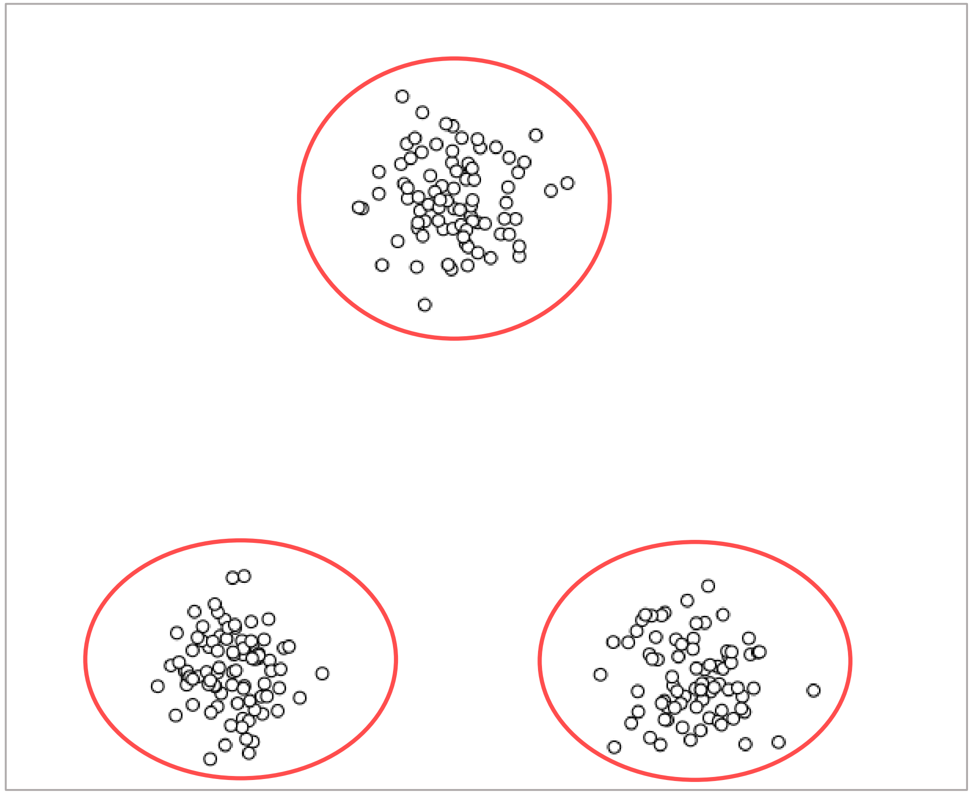
At a data level – knowing which row in the data belongs to which group

Retail store data

P.Whisky	Revenue	Label
0.40	80	1
0.20	60	2
0.21	40	2
0.22	90	3
0.45	75	1

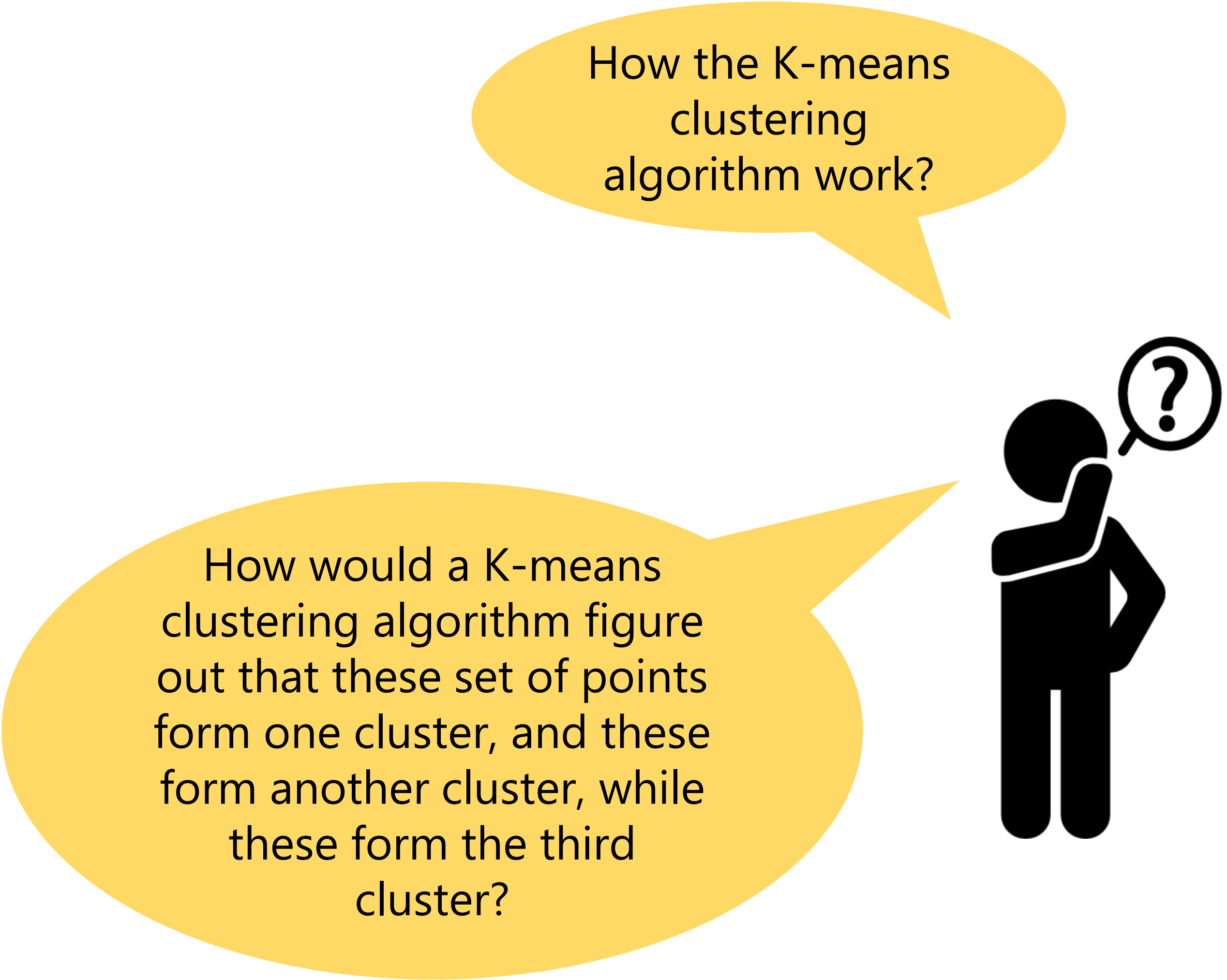
Creation of this label column is one of the potential outputs of a clustering algorithm

Clustering – K-means



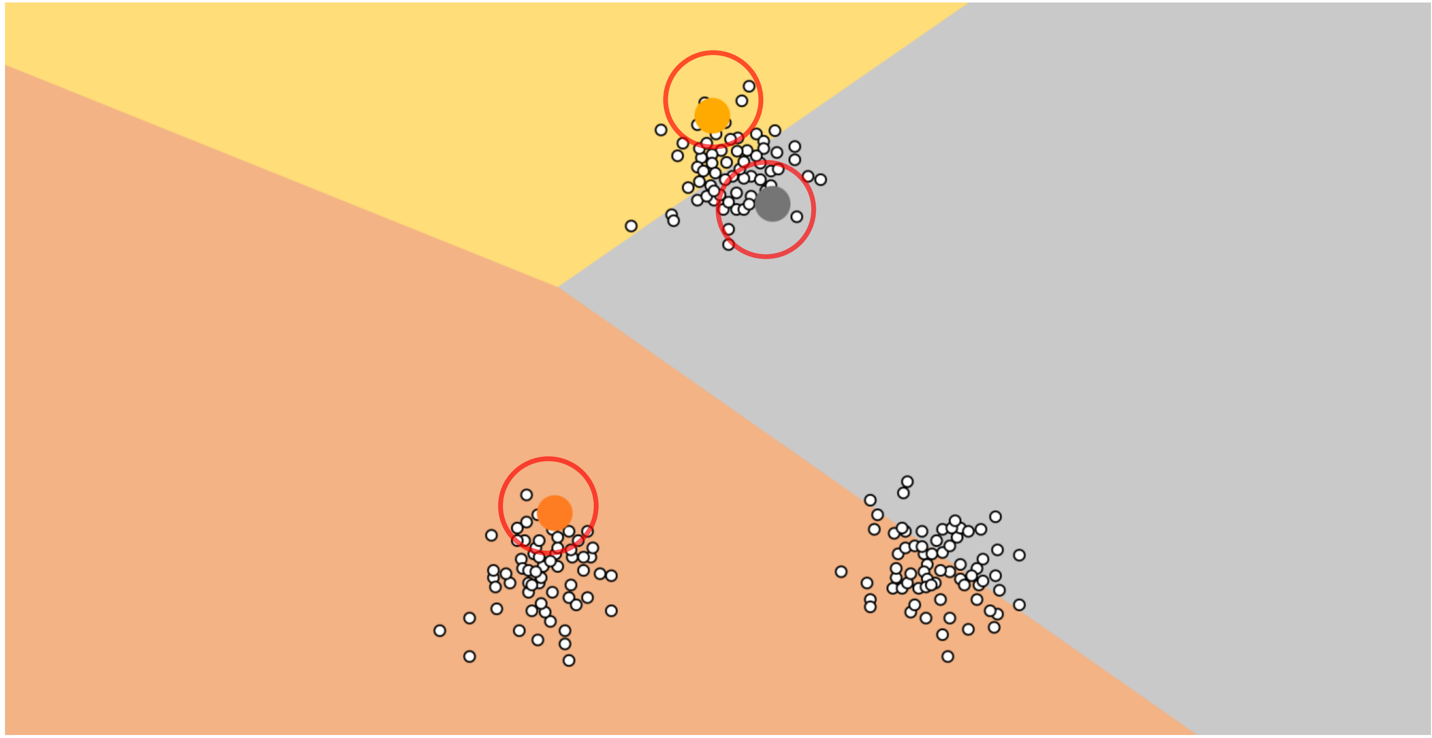
A dataset

We plotted the dataset



Clustering – K-means

K-means clustering algorithm



Assign 3 points as the centres of 3 different clusters

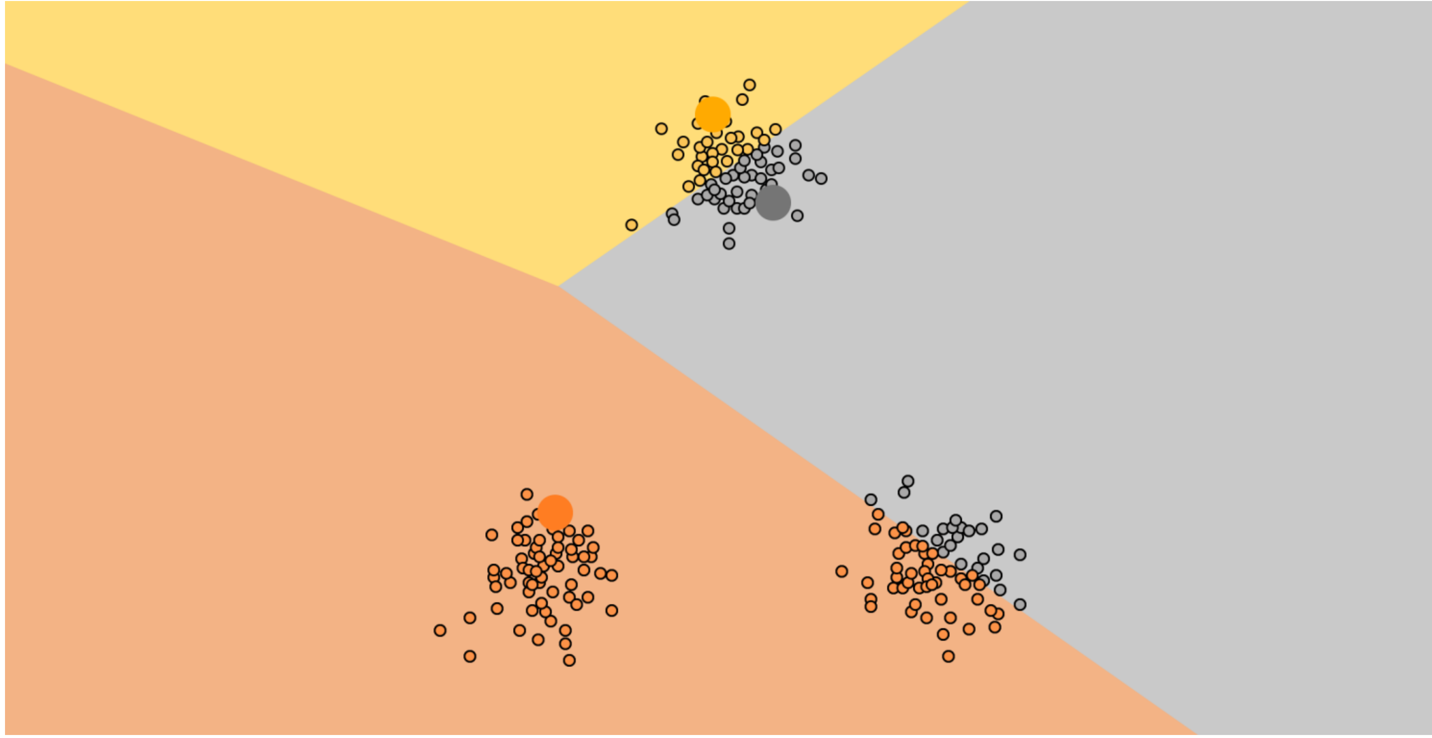
Centre of cluster 1

Centre of cluster 2

Centre of cluster 3

Clustering – K-means

K-means clustering algorithm



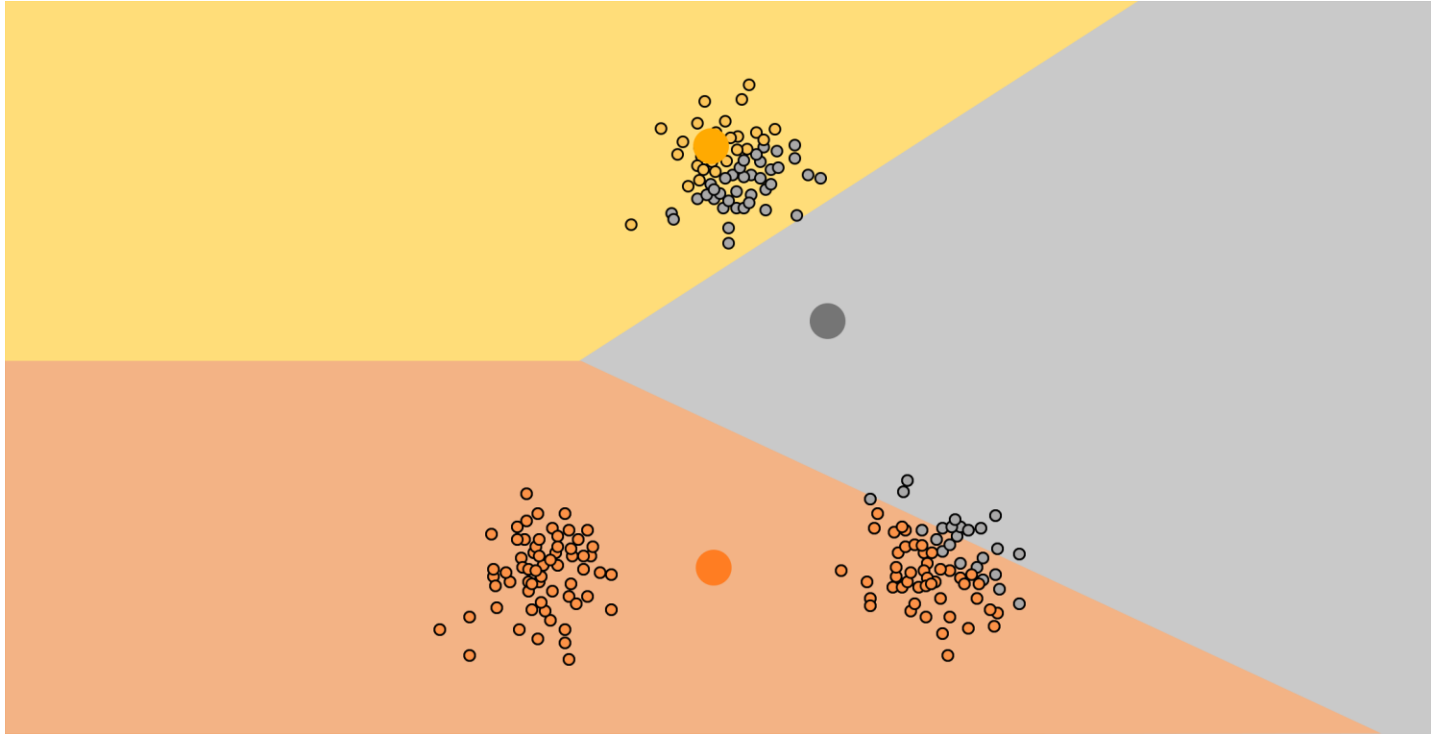
After randomly assigning 3 points as cluster centres

Find the distance of each of the point in the data from these centres

Based on the closeness of that point to a particular cluster centre, allotment of cluster points can be determined

Clustering – K-means

K-means clustering algorithm

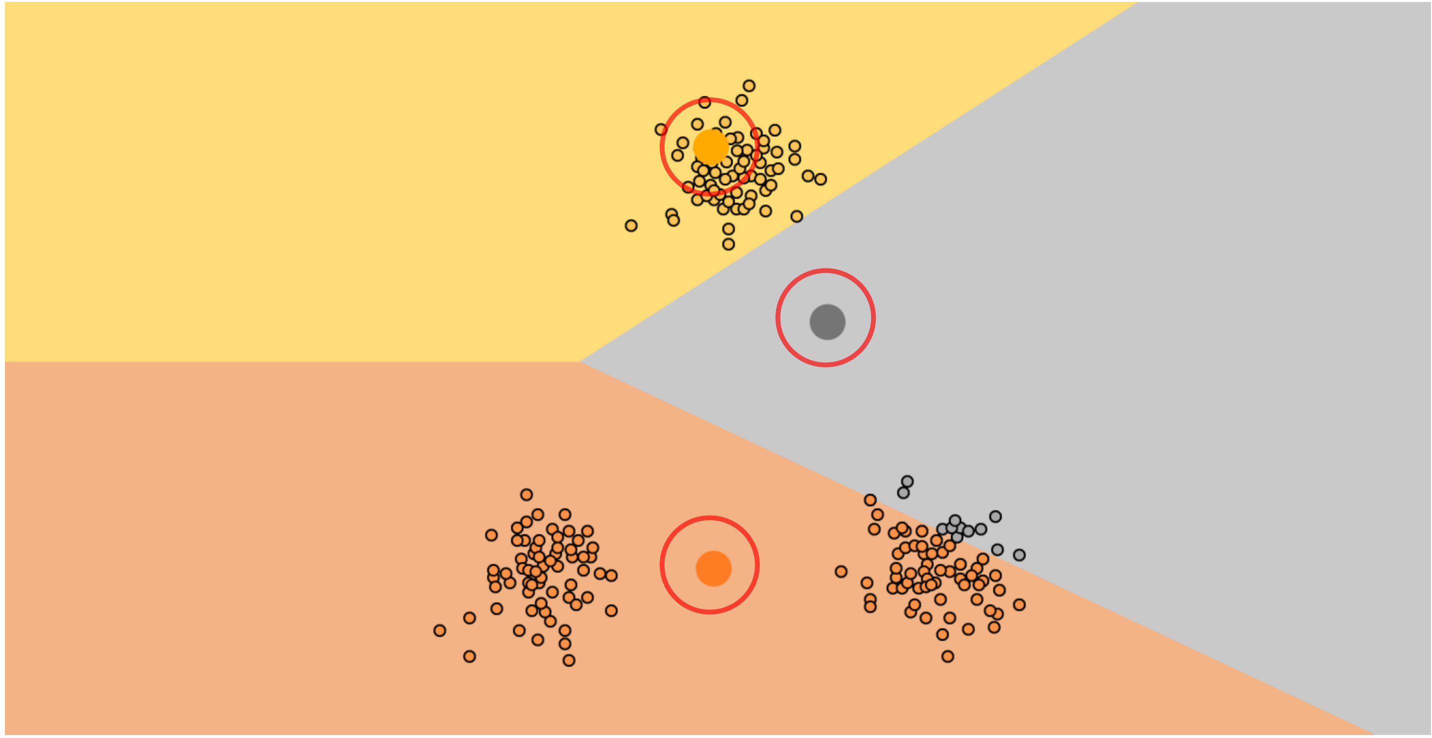


Find the centre of each of these newly formed cluster

Centres of each of these 3 clusters will be re-computed

Clustering – K-means

K-means clustering algorithm



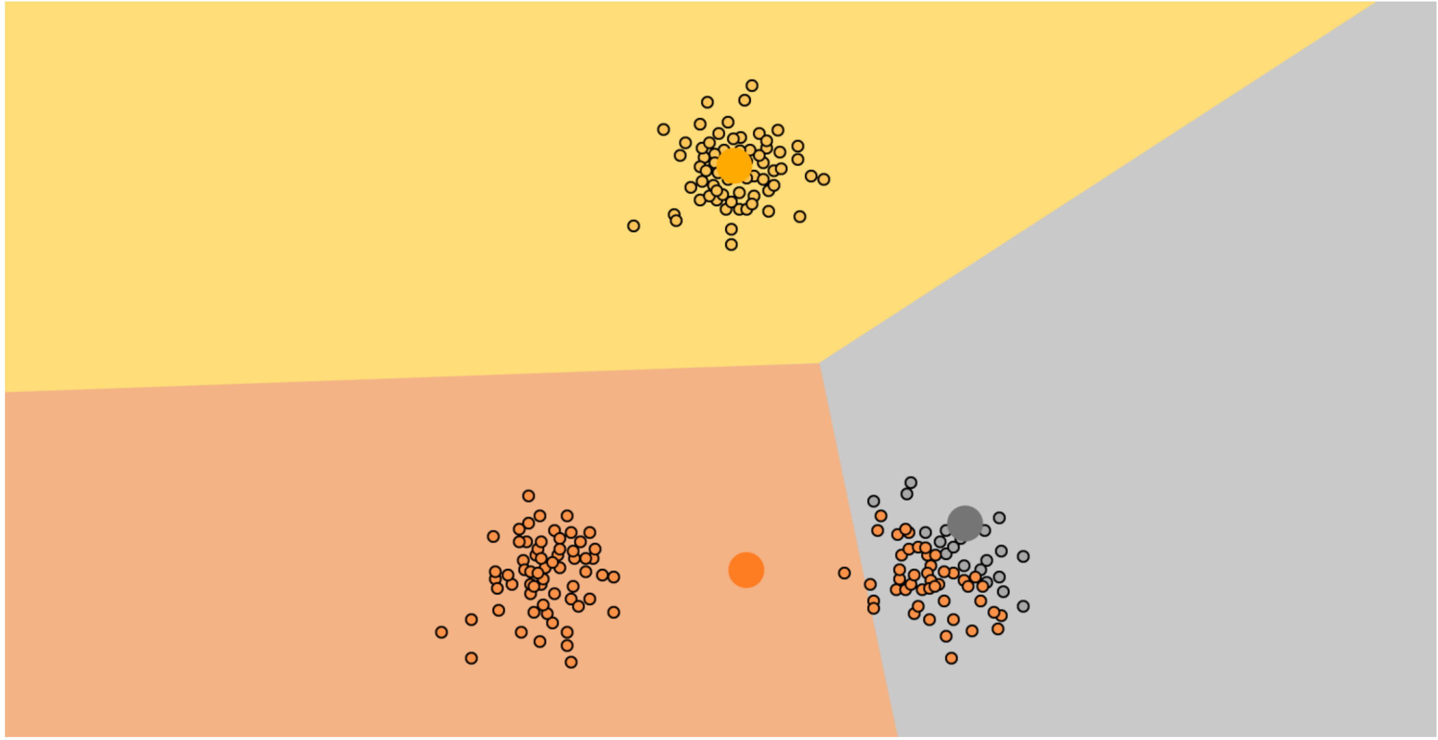
After re-computing the centre

New cluster centres look like

Again, compute the distances of each of the points from these 3 cluster centres and assign cluster labels

Clustering – K-means

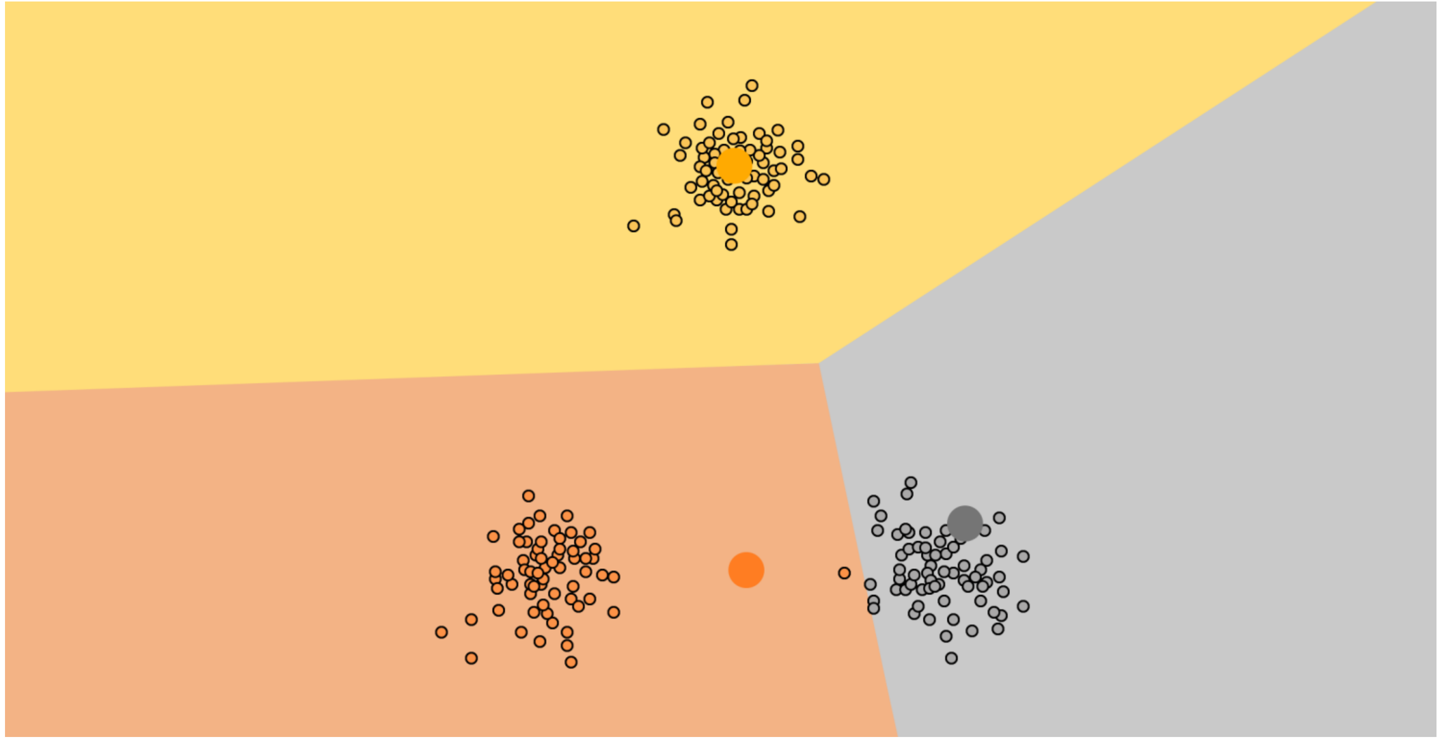
K-means clustering algorithm



Re-compute the cluster centres

Clustering – K-means

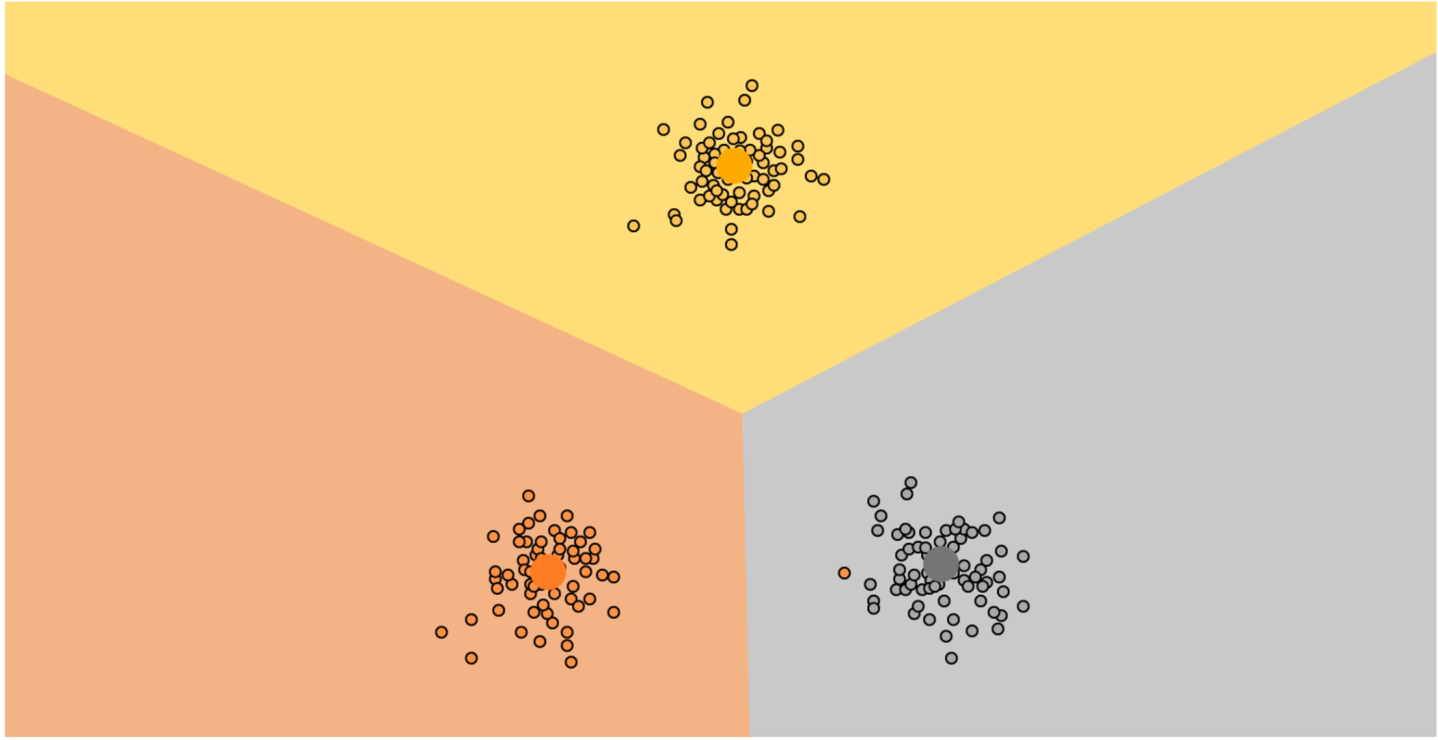
K-means clustering algorithm



Again, find the distance of each point from the 3 cluster centres to decide the allotment of cluster points to clusters

Clustering – K-means

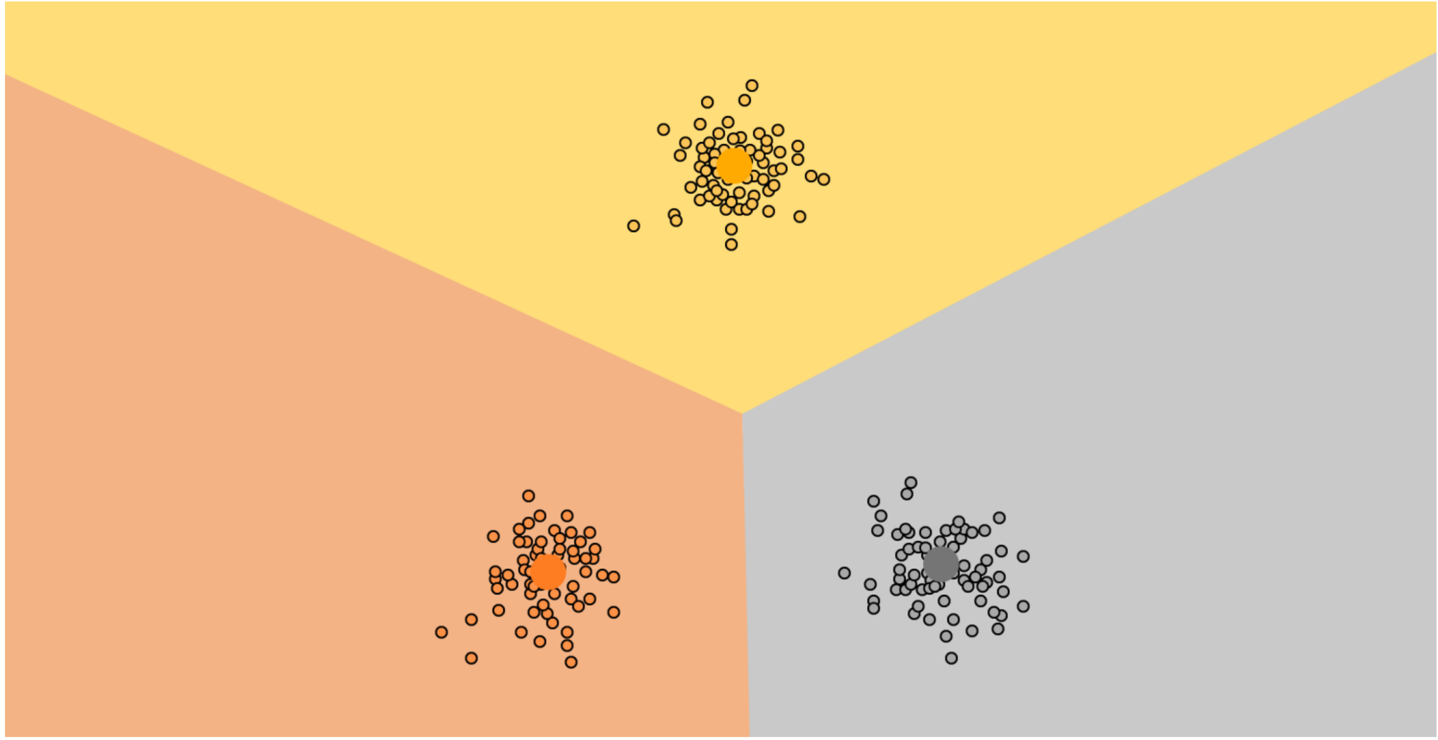
K-means clustering algorithm



Re-compute the cluster centres

Clustering - K-means

K-means clustering algorithm



Again, find the distance of each point from the 3 cluster centres to decide the allotment of cluster points to clusters

Stable solution – a solution where after successive iterations, no point changes its class membership or cluster membership

Clustering – K-means

Randomly assigning k cluster centres

Find the distance of each point in the data from the k cluster centres

Assign cluster labels to each data point based on the proximity of each of the points from the k cluster centres

New clusters are formed

Re-compute cluster centres

Repeat everything till a stable solution is reached

When no data point changes the cluster membership in succeeding iteration

Recap

1. Unsupervised Learning
2. Clustering
3. Clustering - K-means

Class

Clustering

Topic

K-means: Scaling Data, Choosing Number of Clusters and Cluster profiles

K-means

Important points to note while using K-means algorithm

Data used to build K-means models should always be

a. Numeric

b. Be on same scale

How many cluster a data has?

So far, figuring out the number of clusters of data was straight forward as the data could be plotted to decide on the number of clusters

Real life scenario → Knowing the number of clusters a data has is a little involved process

K-means: Numeric Data

In clustering all the data should be numeric

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

Convert categorical data into numeric data

Revenue	Size (Sq. ft.)	Foot fall	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

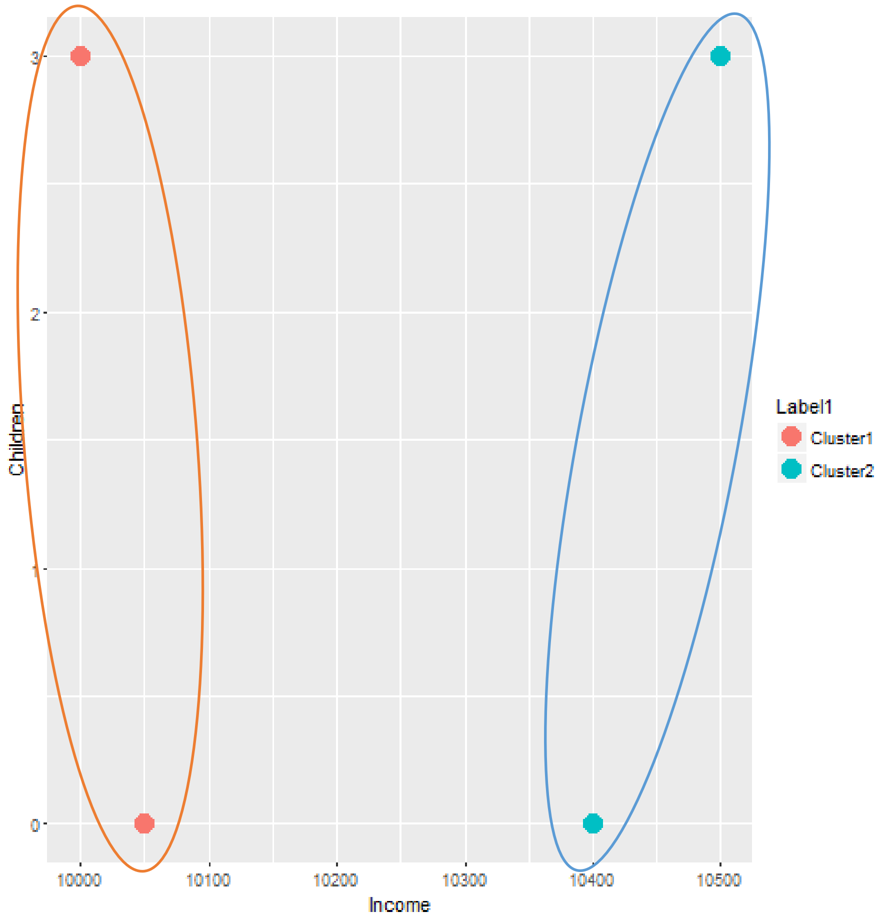
Categorical Variables

Dummy variables

One-hot Encoding — Categorical data

K-means: Scaled Data

In clustering all the variables in a data should be on the same scale



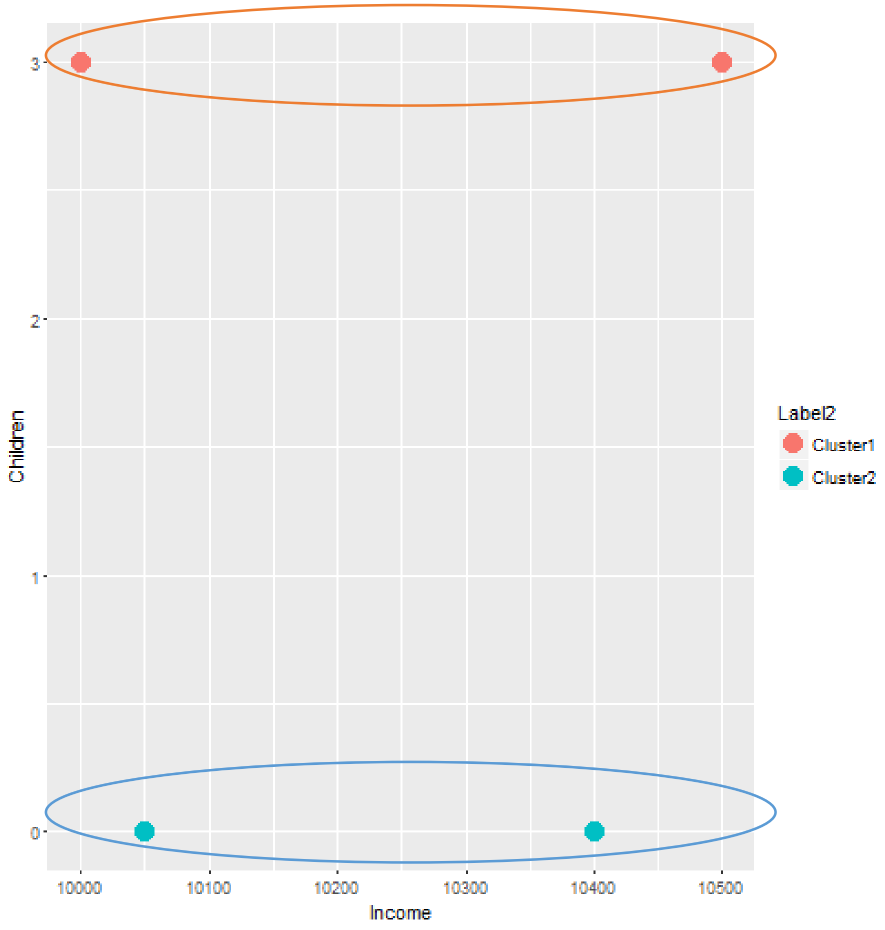
Income	Children
10050	0
10050	2
10400	0
10500	2

Theoretical minimum and maximum for the variable **'Children'** is very different from that of the variable **'Income'**

Both these variables are not on the same scale

K-means: Scaled Data

In clustering all the variables in a data should be on the same scale



Income	Children
10050	0
10050	2
10400	0
10500	2

Both these variables are not on the same scale

K-means: Scaled Data

In clustering all the variables in a data should be on the same scale

Income	Children
-0.98	-1
-0.98	1
0.73	-1
1.23	1

Z transform

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

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

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

Income	Children
10050	0
10050	2
10400	0
10500	2

It is imperative to bring the data on the same scale before a data is fed to a clustering algorithm

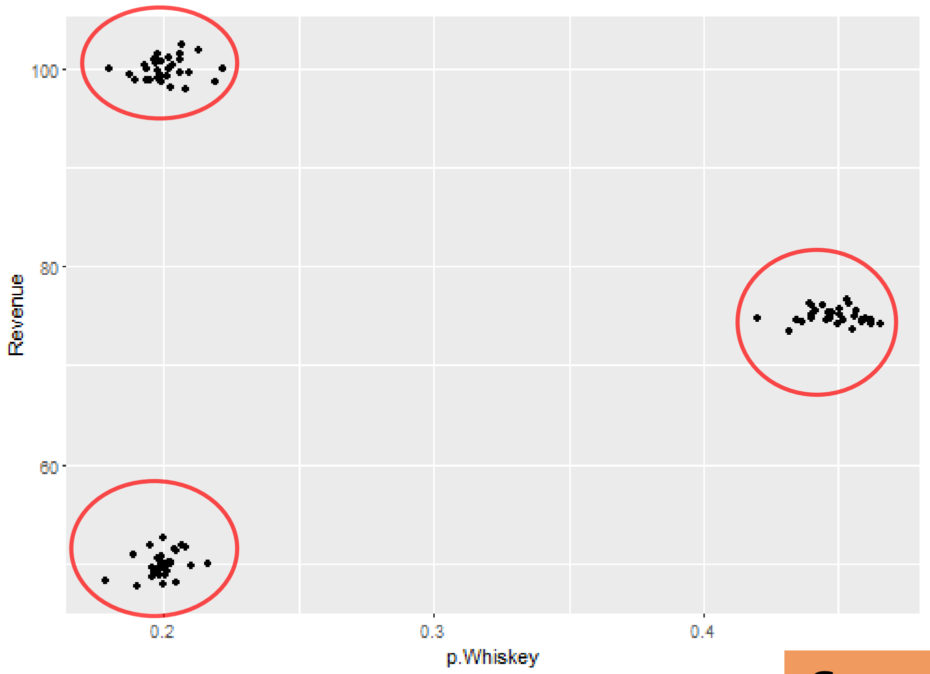
Couple of ways to scale the data

Compute a z transform

Not much difference between the magnitudes of the values of **Income** and **Children,** once the data has been scaled

Z transform - a mean and standard deviation of income column in the data

K-means: Value of K



Retail store data

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

For data that can be visualized, it is possible to guess the approximate number of clusters

How many cluster should be considered in the data?

K-means: Value of K

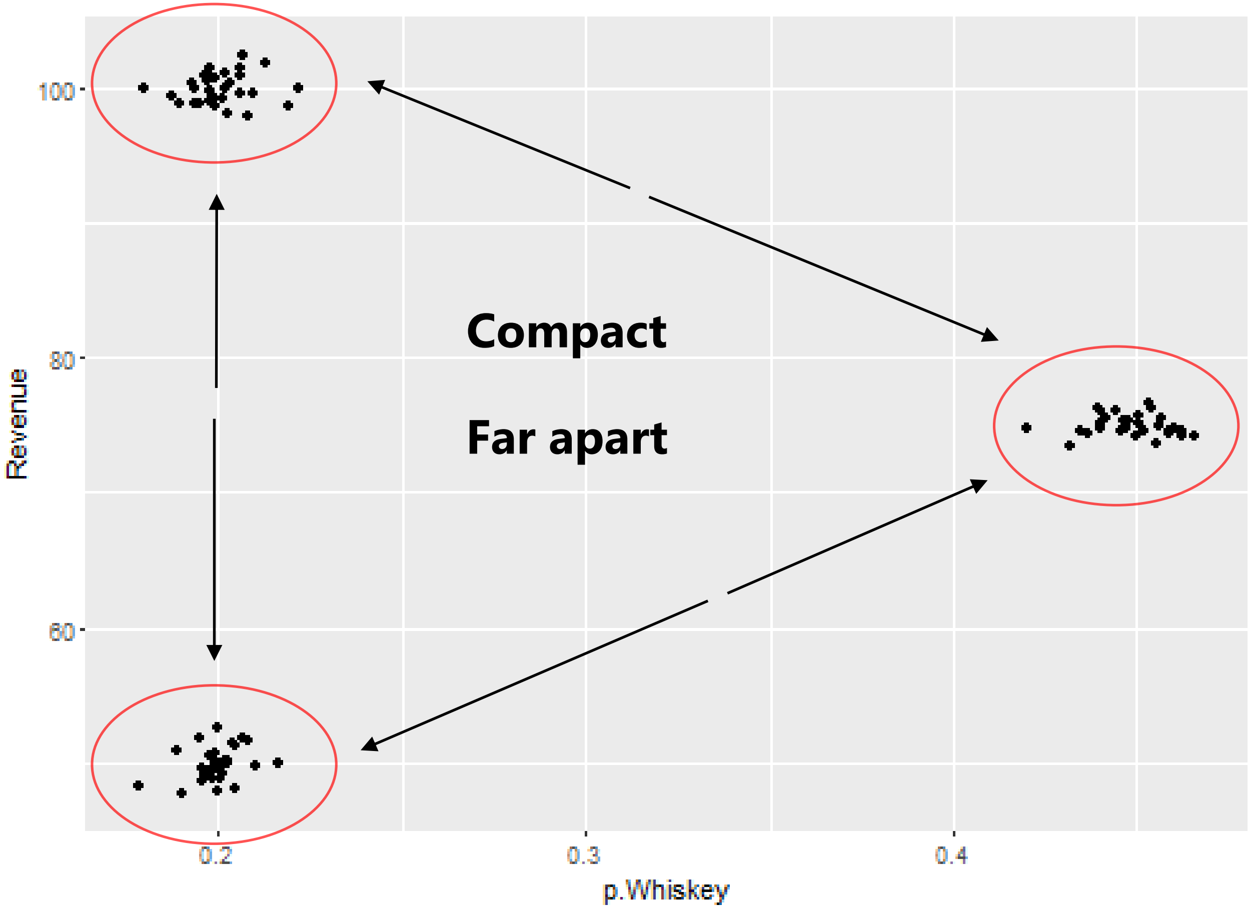
How many cluster should be considered in the data?

Income	Credit Limit	Withdrawals	Card Usage	FICO	Age
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

Sometimes if the **specifics** and **context** of the business problem to solve are known, then it is possible to figure out the numbers of clusters to consider

The marketing team for this bank may just want to look at 3 segments in the data

K-means: Value of K



Retail store data

How to determine the number of clusters in scenarios where the context determining the number of clusters to be considered is absent?

3 clusters

Appropriate number of clusters for this data

Optimum number

Compact and well separated clusters

K-means: Value of K

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

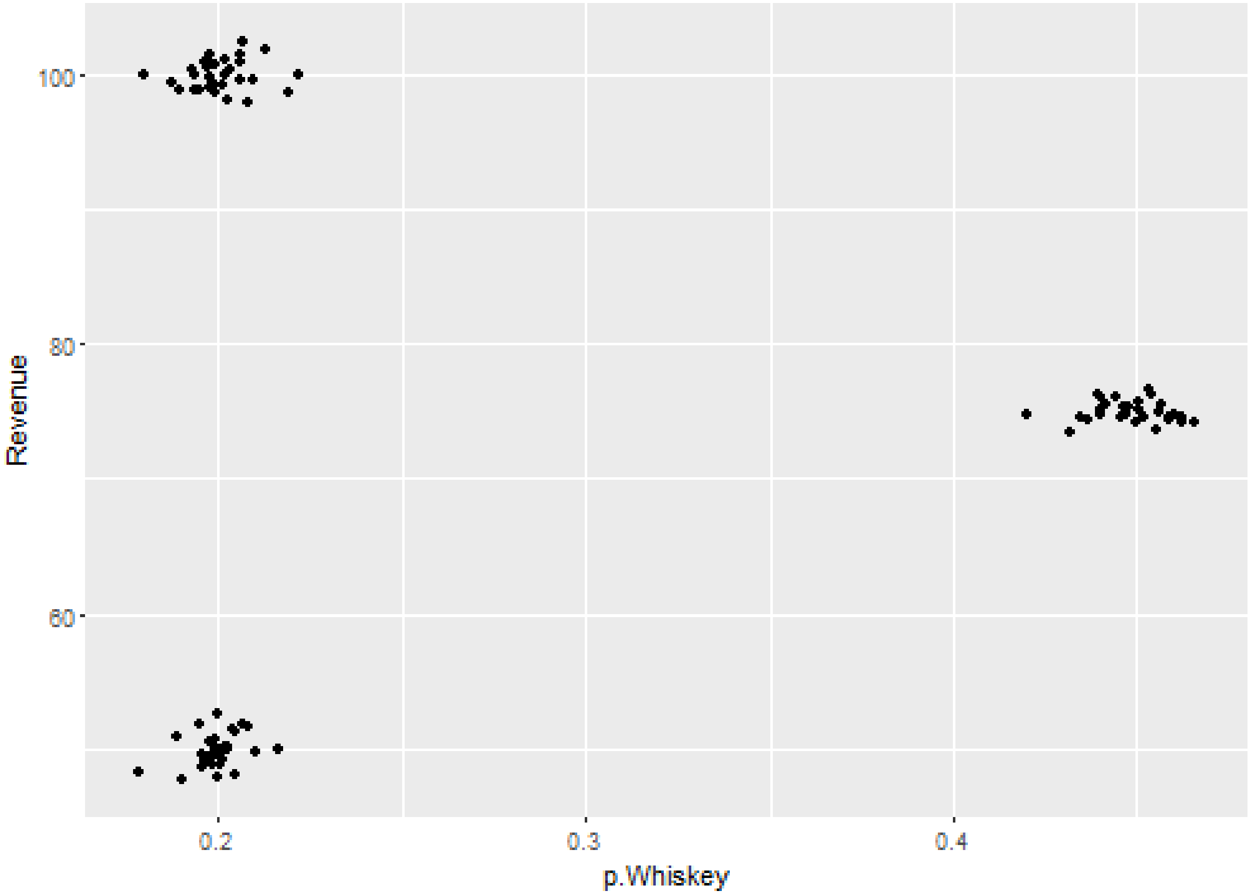
→ **K = 3**

An algorithmic way to find the optimum number of clusters

Calculate some measure of average cluster compactness for a series of values of clusters

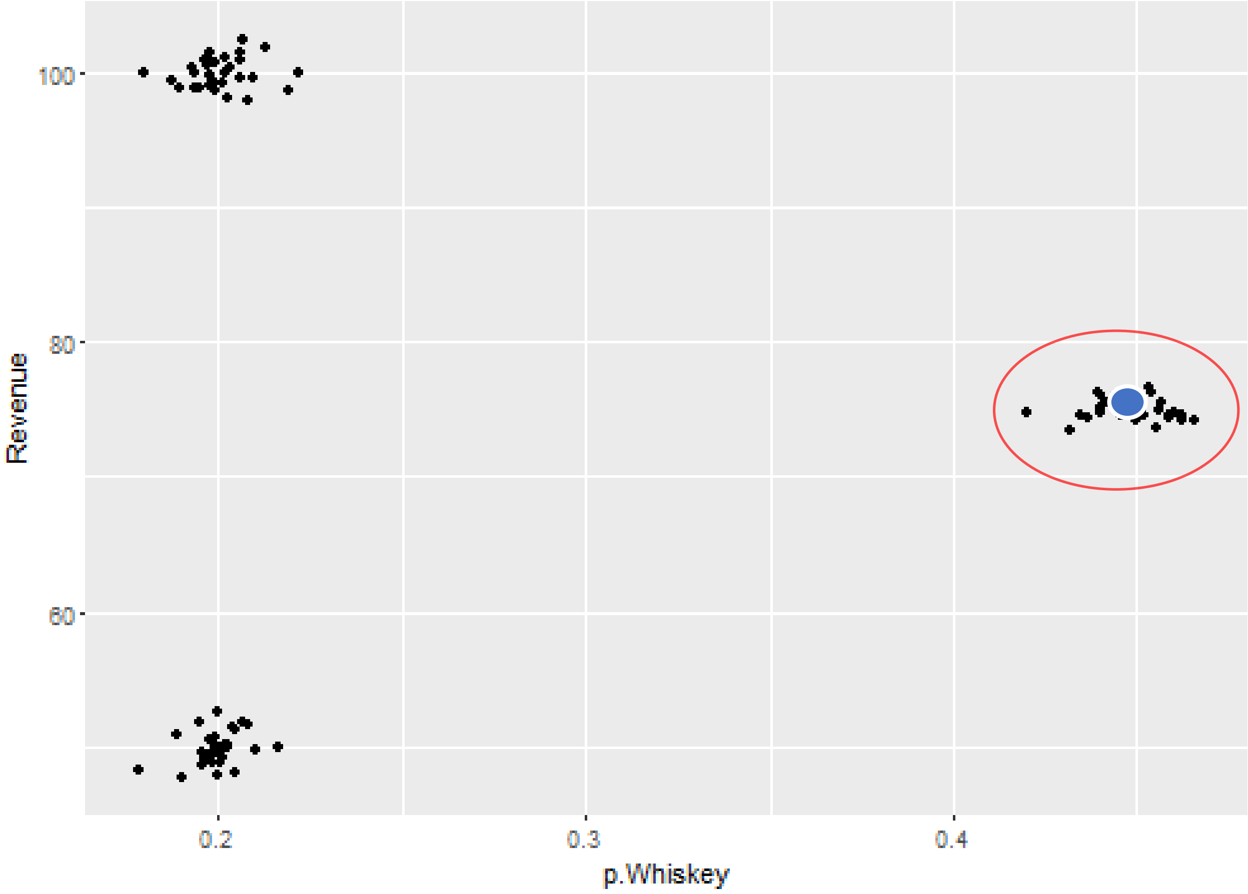
For 3 clusters → The measure of compactness obtains a minima or becomes asymptotic

K-means: Compactness of Clusters



Retail store data

K-means: Compactness of Clusters

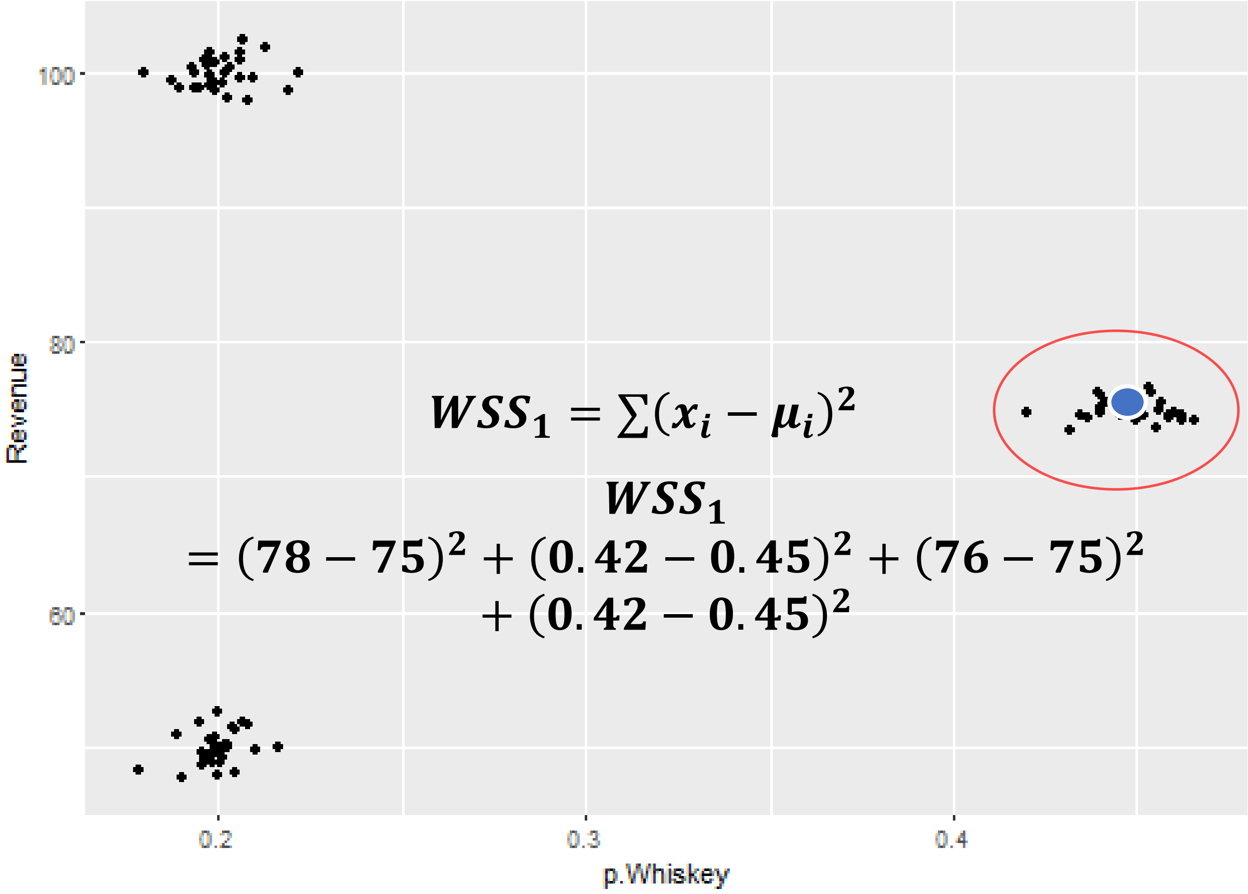


Retail store data

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

Cluster centre

K-means: Compactness of Clusters



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

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

Retail store data

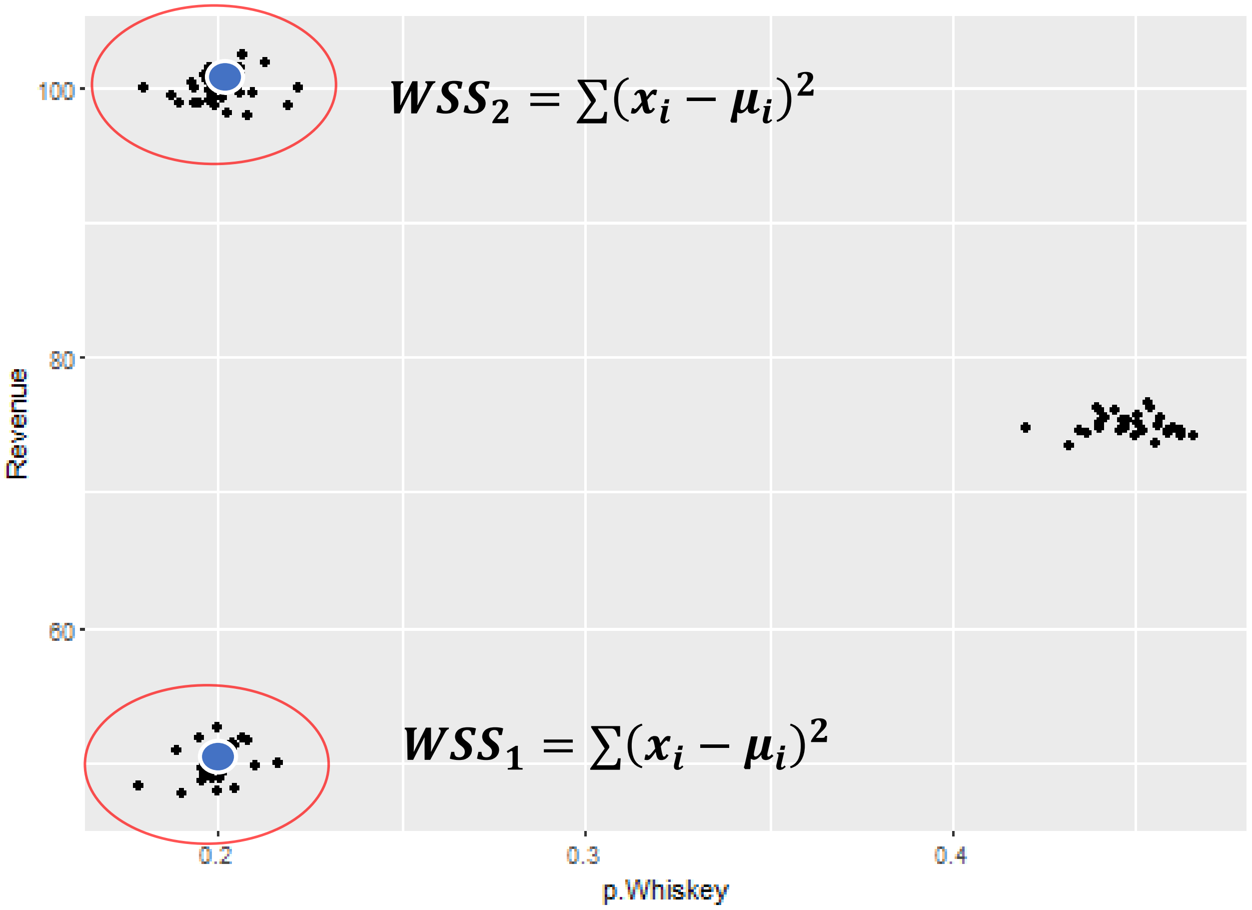
Revenue	P.Whisky
..	..
78	0.42
75	0.45
76	0.42
..	..
..	..
..	..

Other points belonging to the cluster

Computing the total squared distance of each point of this cluster from its cluster centre

WSS or **Within Sum of Squares** – measure the compactness of the cluster

K-means: Compactness of Clusters



Retail store data

Arrive at a consolidated measure of cluster compactness

$$WSS_{Total} = WSS_1 + WSS_2 + WSS_3$$

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

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

Each value of K is computed

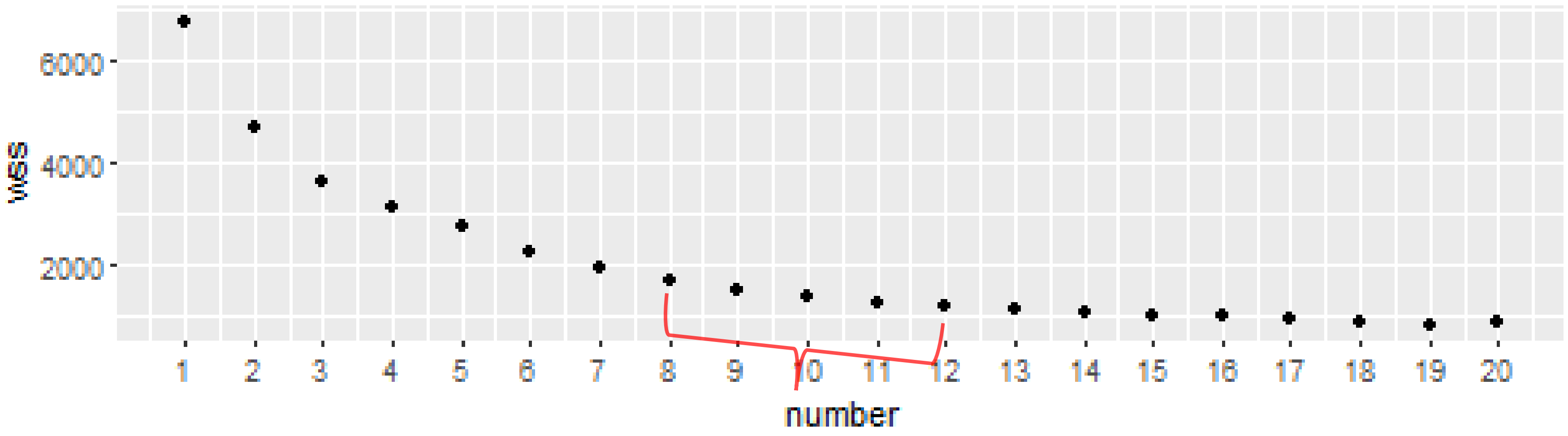
A decision is taken

K-means: Scree Plot/ Elbow Plot

Scree Plot or **Elbow plot** – when the total within SS for each value of K is plotted to get an idea on the optimum number of clusters

K-means: Screen Plot/ Elbow Plot

Scree plot or Elbow plot



Decrease in WSS is not substantial after around 8 clusters

Improvement in cluster compactness is very minimal after cluster 8

Improvement flattens out at around 12 clusters

8 to 12 clusters can be the **optimum number of clusters**

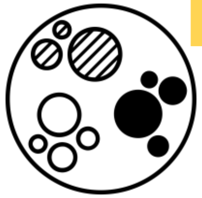
There is some subjectivity of the analyst in determining the ball park number of optimal clusters

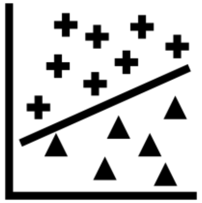
K-means: Creating Cluster Profile

Next step after finding the range of optimal clusters

Create cluster profiles

Creating cluster profiles - understanding what each cluster represents after cluster models have been created

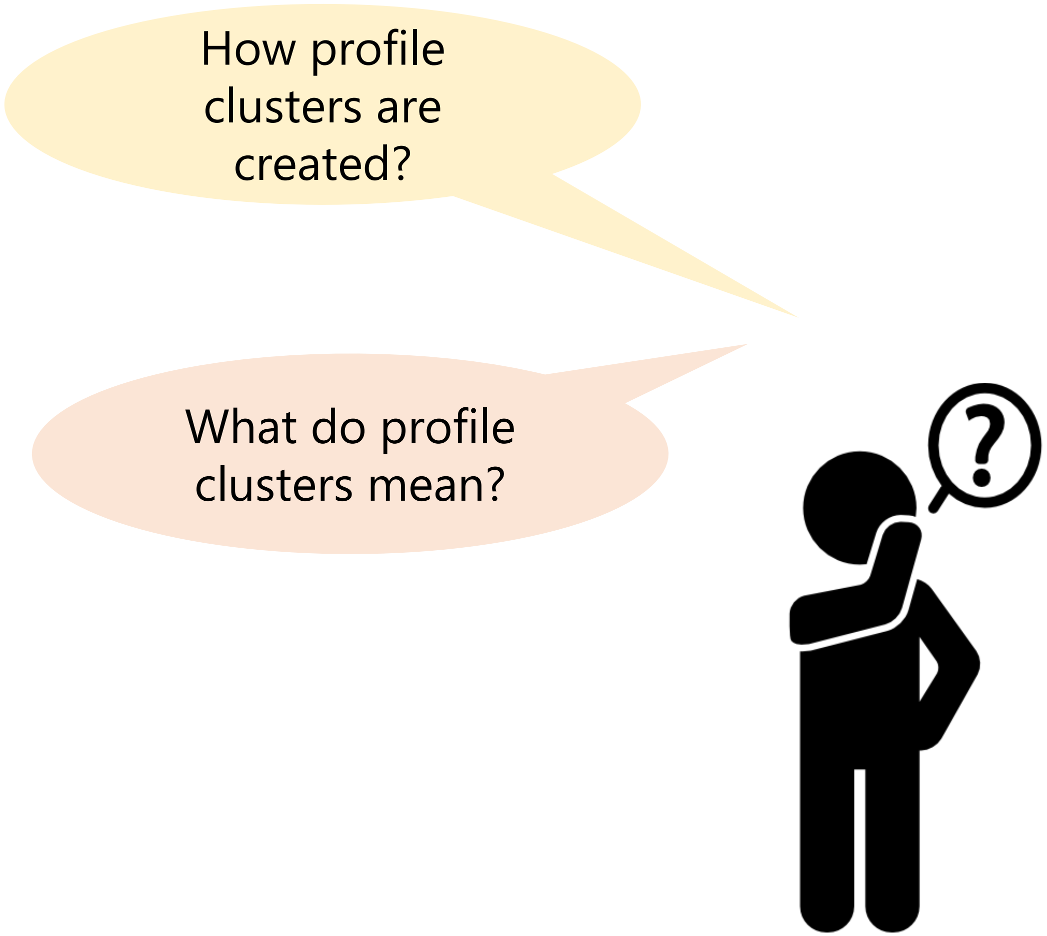




Based cluster profiles the number of the cluster model is figured as suggested by the Elbow curve

Example – deciding if 8 cluster model is better or 11 cluster model

K-means: Creating Cluster Profile



K-means: Creating Cluster Profile

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

Which row in the data stands for which cluster is known

Average value of the variables in the data itself can also be exploited

Global means

This can help in deciding whether the clusters, that have been created, are meaningful from a business point of view

K-means: Creating Cluster Profile

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

Clusters will be more meaningful if they are different

Global average

	Cluster 1	Cluster 2	Cluster 3
Average revenue	High	Near global average	Low
Percentage of whiskey sales	High	Near global average	Low

K-means: Creating Cluster Profile

Computing z values to do a comparison with global mean

Z values with high positive magnitudes - cluster means are larger than the global means

Z values with negative magnitude - cluster means are smaller than the global means

Variable profiling helps in understanding whether the clusters created are meaningful

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

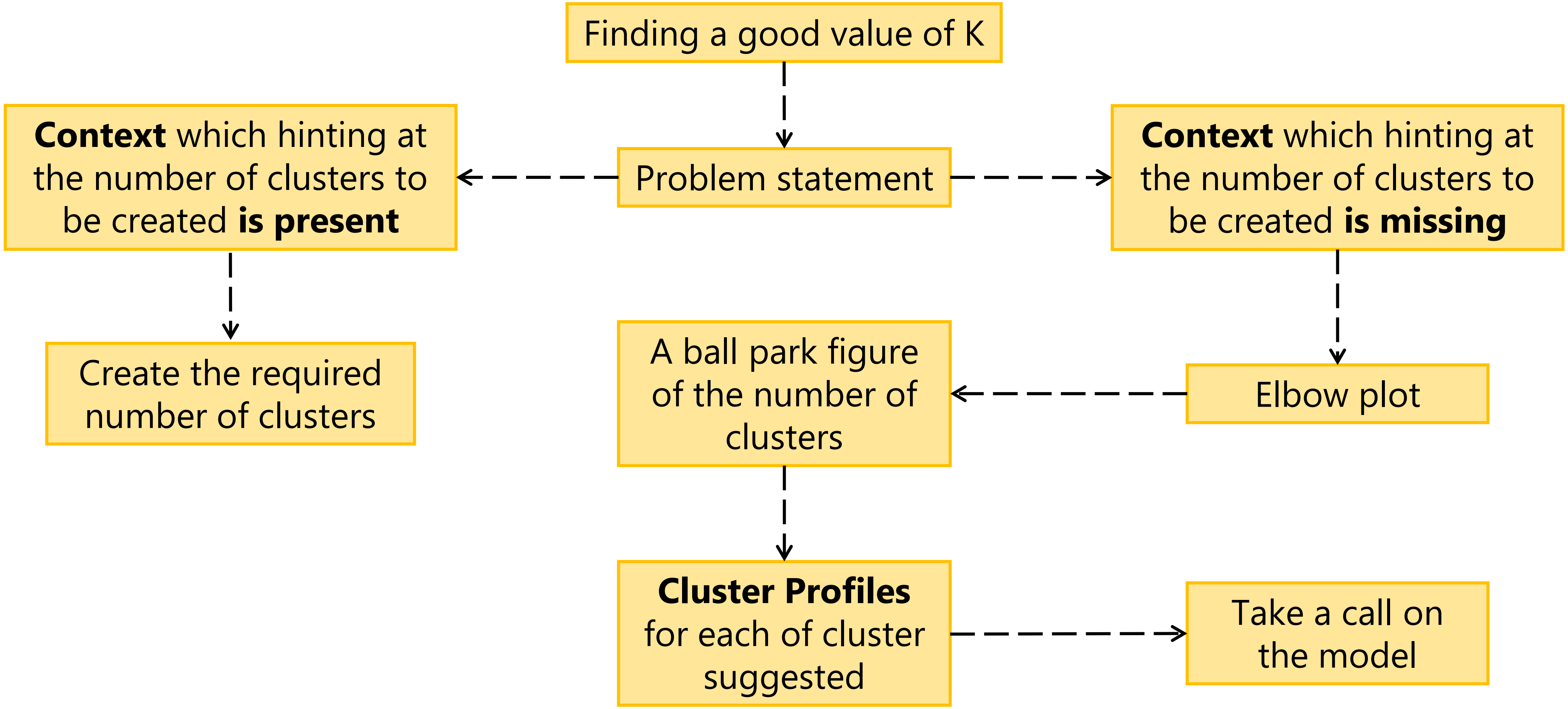
For uniform comparison of differences

$$Z = (x - \mu)/\sigma$$

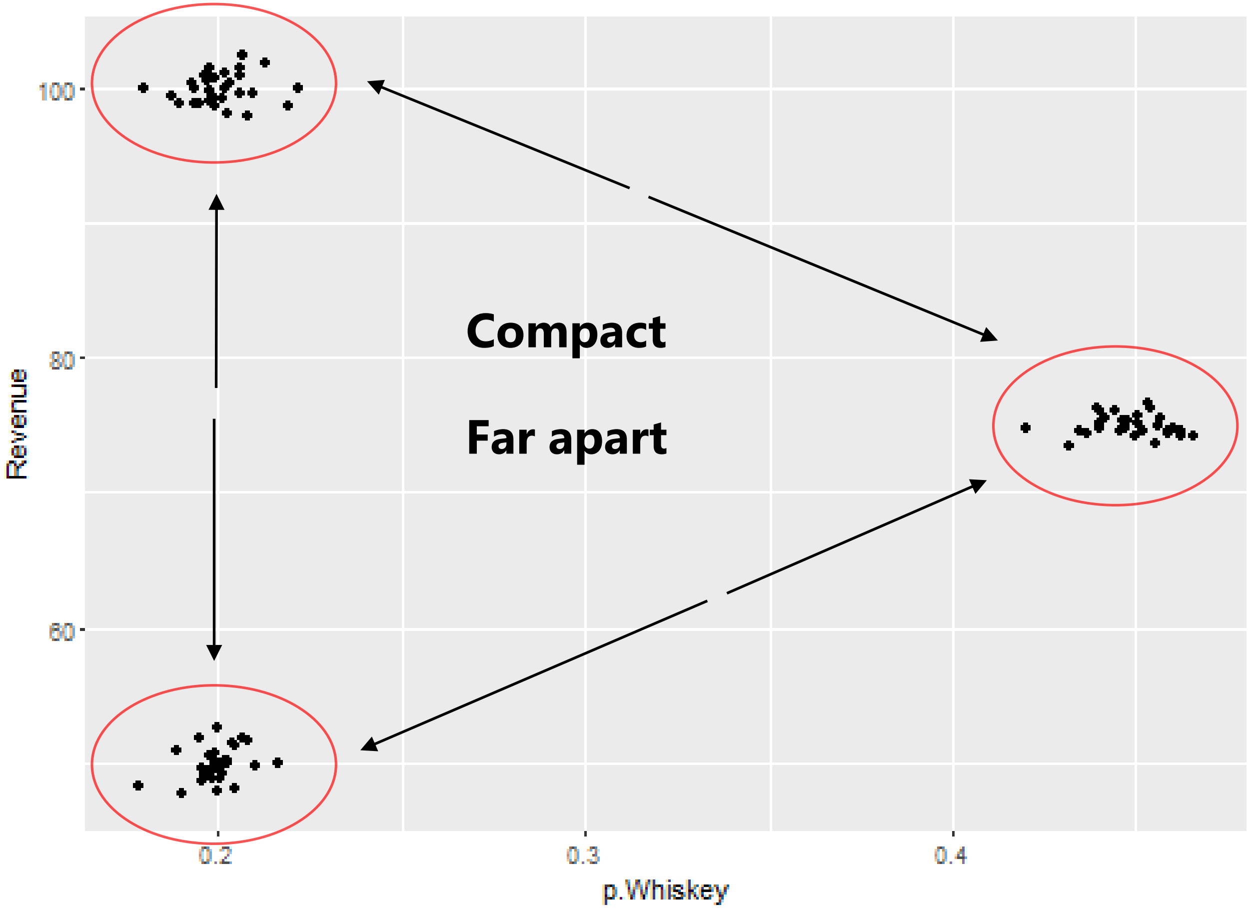
Mean Revenue= 120, std = 10
Z Revenue Cluster 1 = 8
Z Revenue Cluster 2 = 0.5
Z Revenue Cluster 3 = -4.5

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

K-means: Value of K



K-means: Silhouette Measure



Retail store data

Finding the value of K

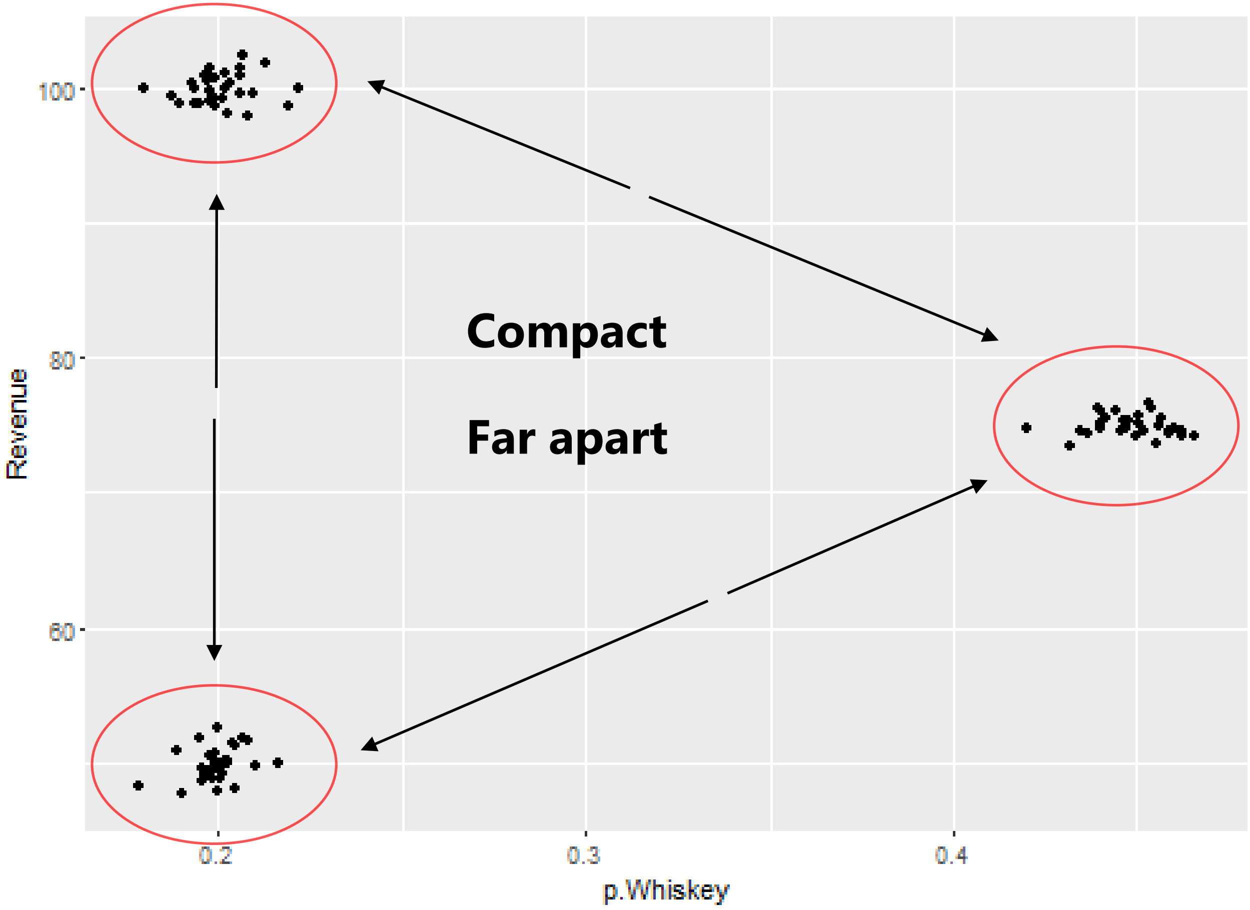
Elbow plot

Silhouette measure

A number that lies between -1 to 1

A higher value represents a good fit of the points given their cluster assignments

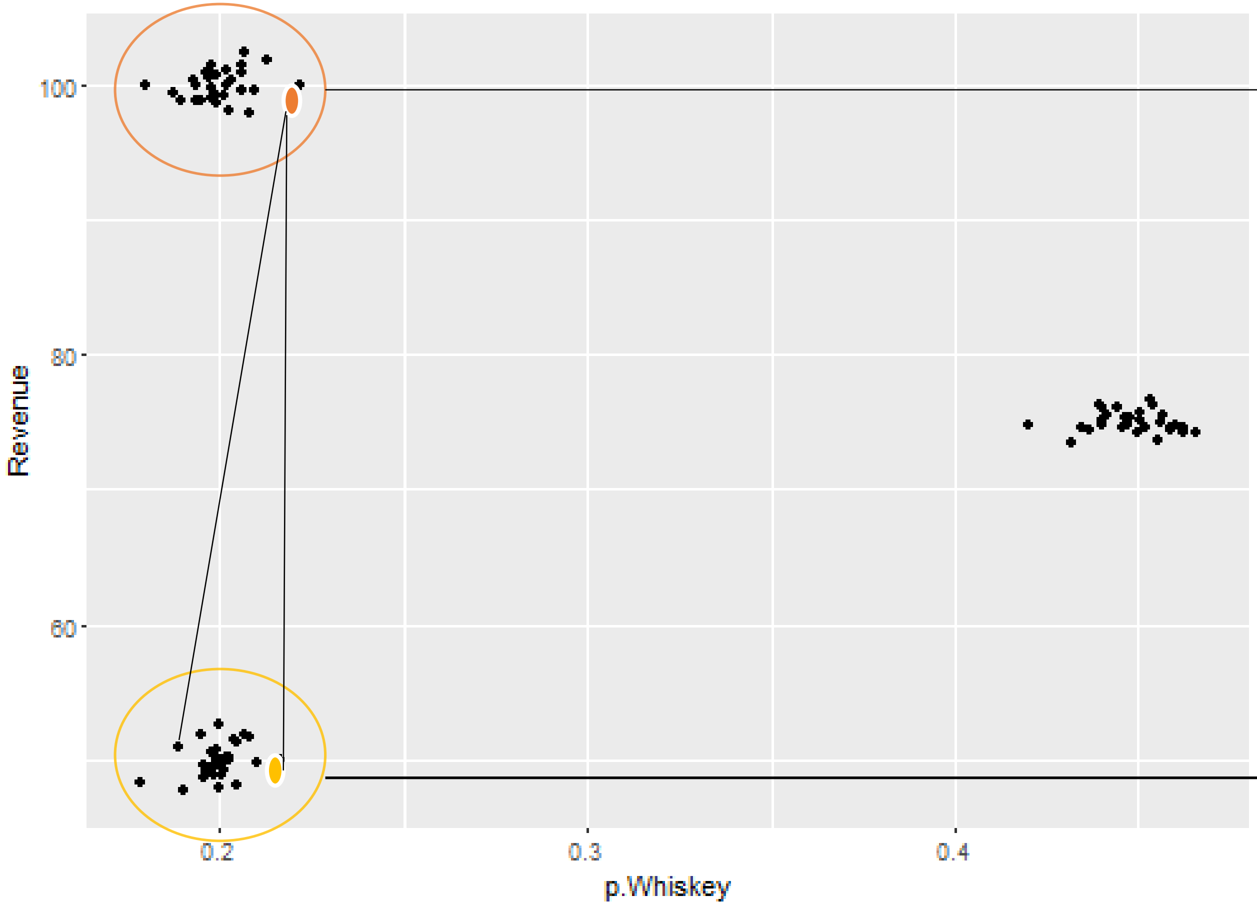
K-means: Silhouette Measure



Retail store data

Silhouette Measure – a measure used to calculate the notions of clusters being far apart and compact

K-means: Silhouette Measure



Retail store data

Computing Silhouette Measure

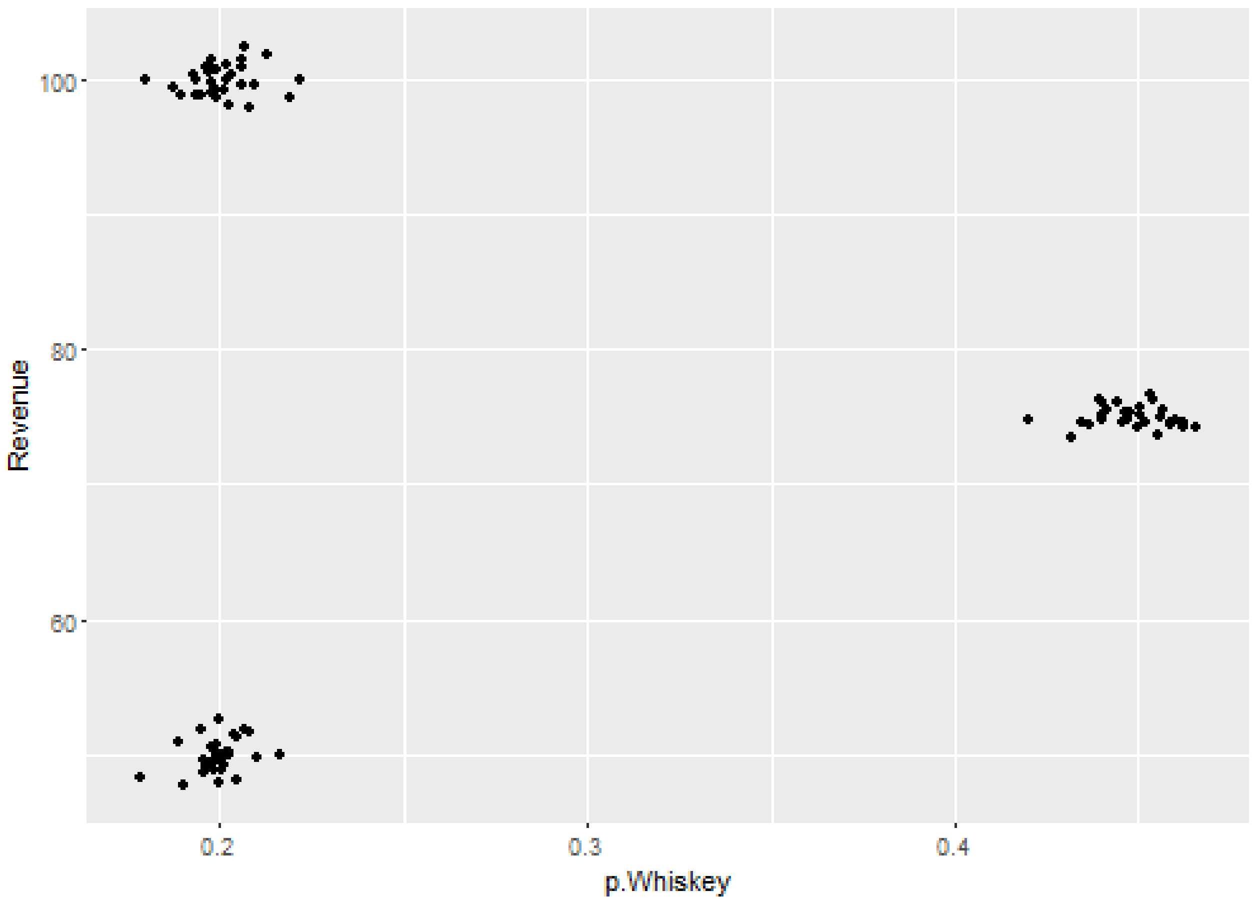
a: Mean distance of a sample from all the points in its **respective cluster**

a - measures the compactness of the cluster, and ideally value is small

b - measures the separation of clusters, and ideally value is large

b: Mean distance of a sample from all the points in its **nearest cluster**

K-means: Silhouette Measure



Retail store data

Computing Silhouette Measure

$$\boldsymbol{Silhouette} = \frac{(\boldsymbol{b} - \boldsymbol{a})}{\mathbf{max}(\boldsymbol{a}, \boldsymbol{b})}$$

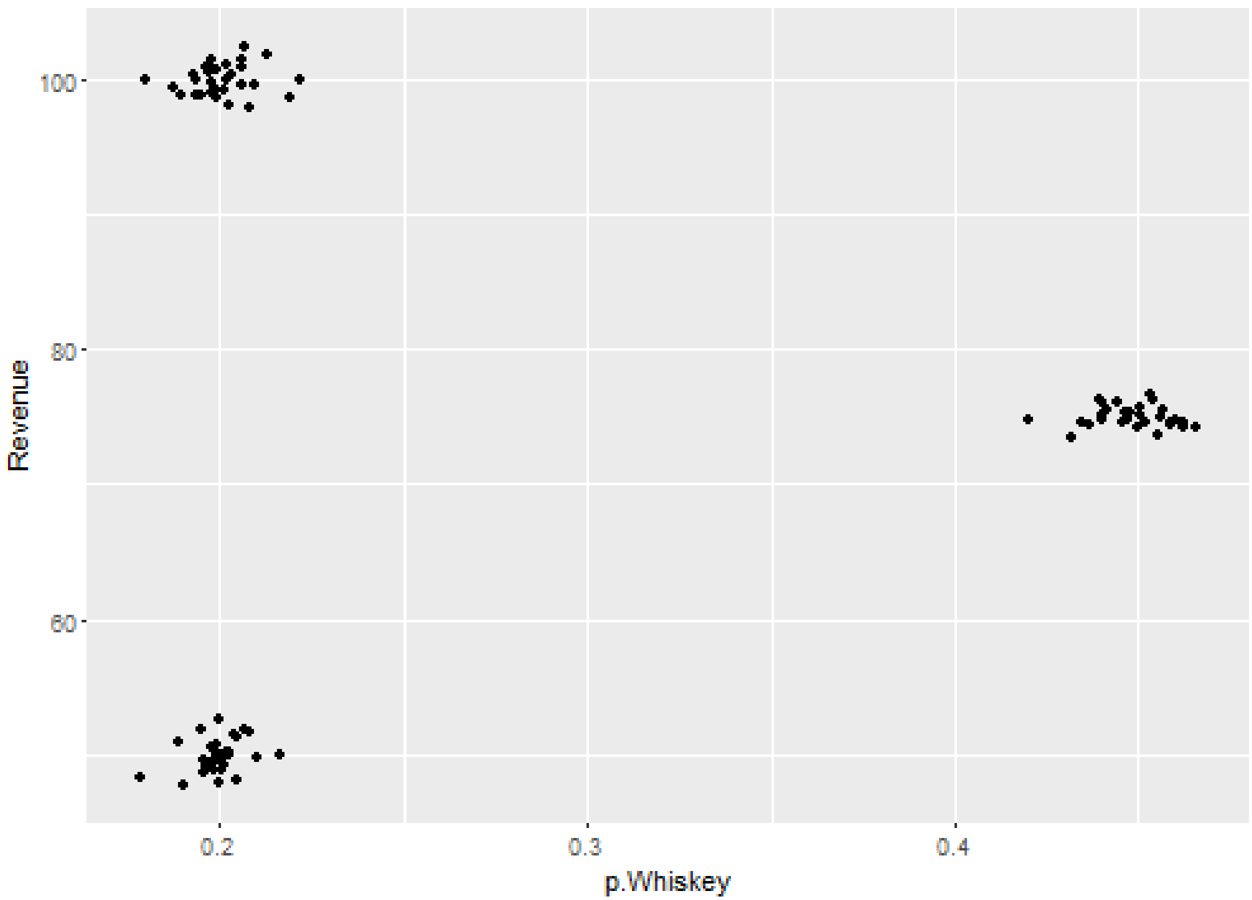
a - measures the compactness of the cluster, and ideally value is small

b - measures the separation of clusters, and ideally value is large

Silhouette value of -1 – **sub-optimal clustering**

Silhouette value of around 1 – **good clustering**

K-means: Silhouette Measure



Retail store data

Computing Silhouette Measure

$$\boldsymbol{Silhouette} = \frac{(\boldsymbol{b} - \boldsymbol{a})}{\mathbf{max}(\boldsymbol{a}, \boldsymbol{b})}$$

Use silhouette measure to calculate the approximate optimal number of clusters

Whatever number of clusters, an average silhouette value closer to 1 is obtained

Those many clusters will be formed

Recap

K-means

1. Numeric Data
2. Value of K
3. Compactness of Clusters
4. Scree Plot/ Elbow Plot
5. Creating Cluster Profiles
6. Silhouette Measure

Class

Clustering

Topic

Agglomerative Clustering

Agglomerative Clustering

Popular algorithm used to create clusters in data

K-means

Agglomerative clustering or Hierarchical clustering

How Agglomerative clustering algorithm works?

Code Demo

Optimum Number of Clusters

How the optimum number of clusters for a hierarchical clustering model is decided?

Business problem

↓

Context giving an implicit or explicit hint about the number of clusters to be used

↓

Decide the number of clusters that are to be extracted out of a hierarchical clustering model

Optimum Number of Clusters

Example: a marketer wants to look at only 4 segments in a given data

Hierarchical clustering → 4 clusters → What they mean

Silhouette measure can be used to calculate optimum numbers of clusters, theoretically or algorithmically

Now at whatever number of clusters, the silhouette measure approaches 1 or is nearer to 1

Hierarchical clustering model

Scrutinizing the cluster profiles by inspecting the clusters created in context of the business

If cluster profiles make a business sense then it is adopted, otherwise the process is repeated

Recap

1. Agglomerative Clustering
2. Code Demo
3. Finding Optimum Number Of Clusters