STATISTICS AND BIG DATA '25-'26

K-means Clustering

Dr. Niccolò Salvini

Adjunct Professor @UCSC campus Rome,
Sr. Data Scientist

MADE WITH
beautiful.ai

— K-means clustering concepts —

1 **principal ideas and overview**

— K-means in R —

2 **minimal R code!**

— live coding session! —

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

Section 1

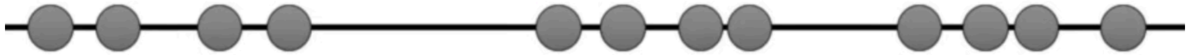
K-means Principles

UNIVERSITÀ
CATTOLICA
del Sacro Cuore

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**



Cluster on a line

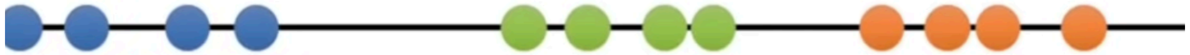
some data on a line...

You may guess some clusters, how would you do taht?

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**

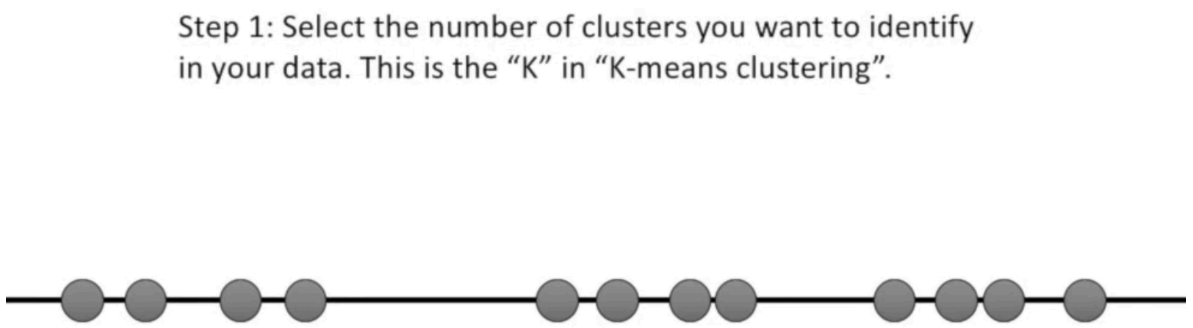


This is how a human would do that

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



Step 1: Select the number of clusters you want to identify in your data. This is the “K” in “K-means clustering”.

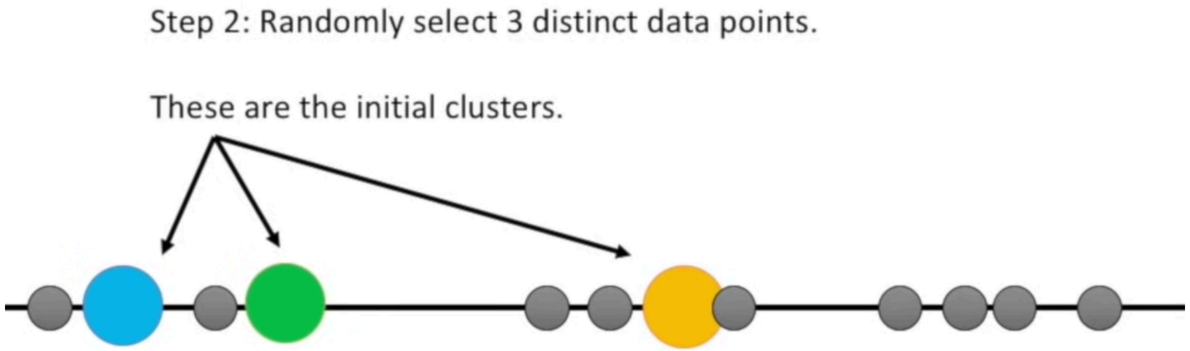
select # k

there a way to do that, we are going to see that later on. For now let's trust our guts feelings

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



Step 2: Randomly select 3 distinct data points.

These are the initial clusters.

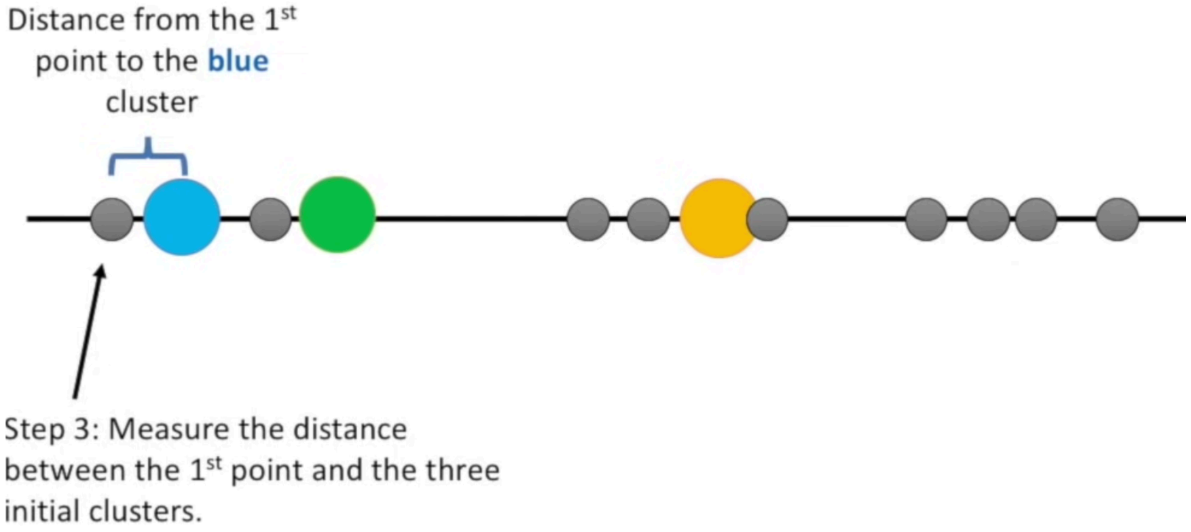
step 1

init algorithm, randomly assign some grey bubbles to clusters.

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



Distance from the 1st point to the **blue** cluster

Step 3: Measure the distance between the 1st point and the three initial clusters.

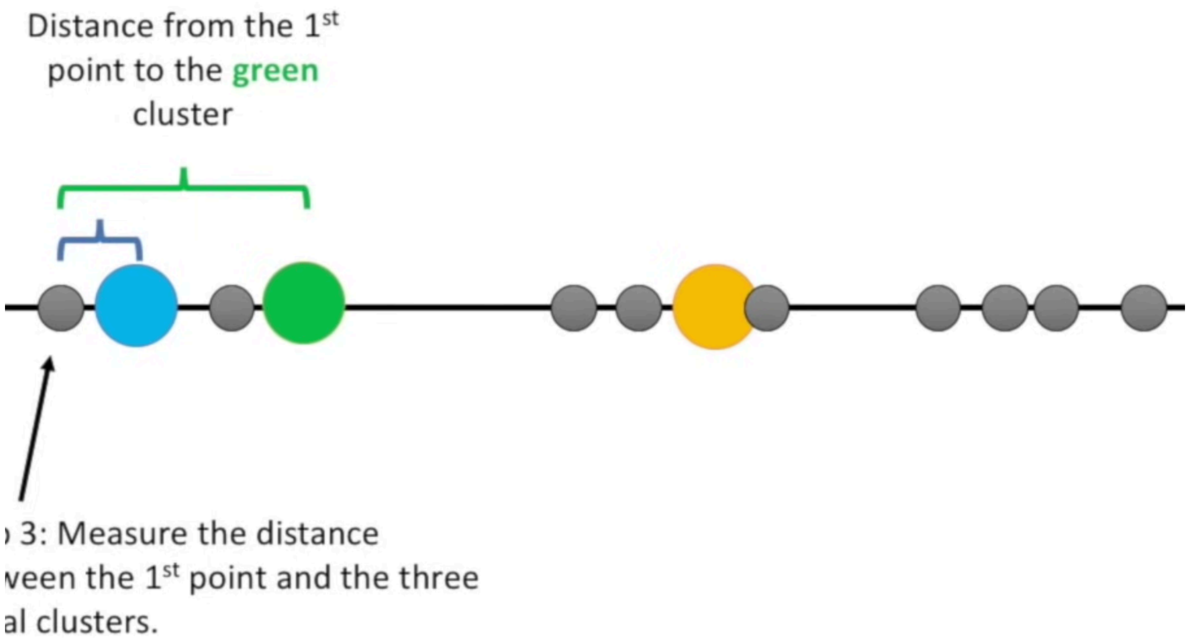
step **2**

compute each distance from first point to each of assigned colored bubble,

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**

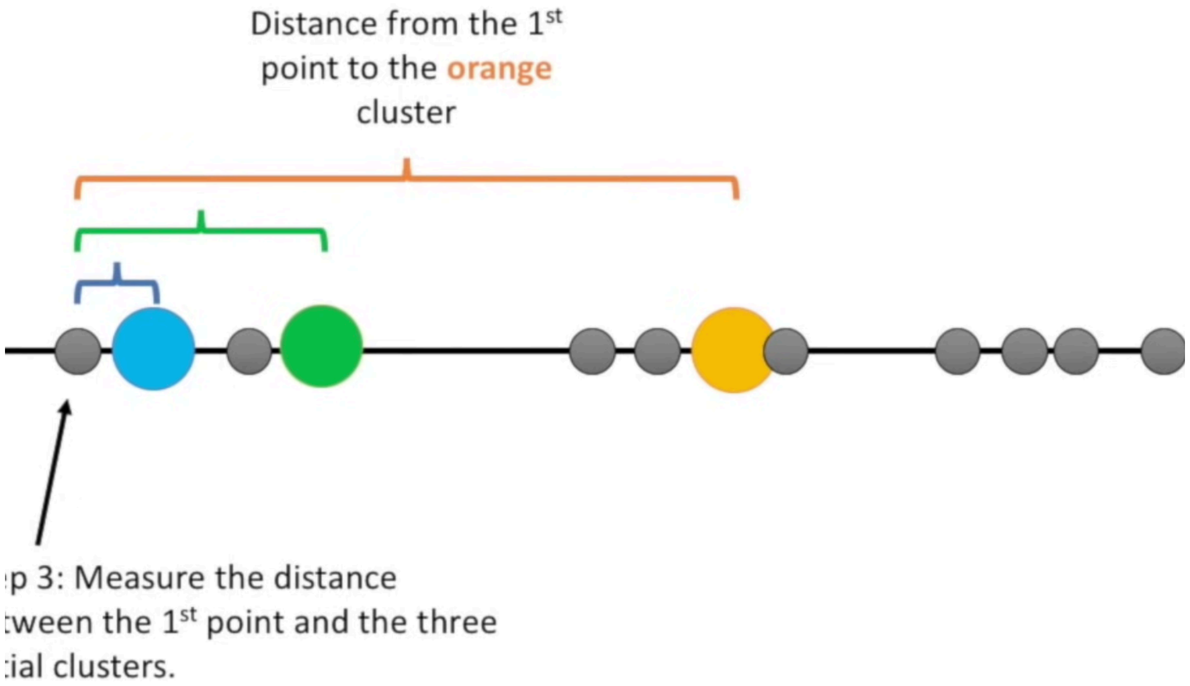


Distance from the 1st point to the **green** cluster

3: Measure the distance veen the 1st point and the three al clusters.

now for **green**

cluster with **heatmaps**



Distance from the 1st point to the orange cluster

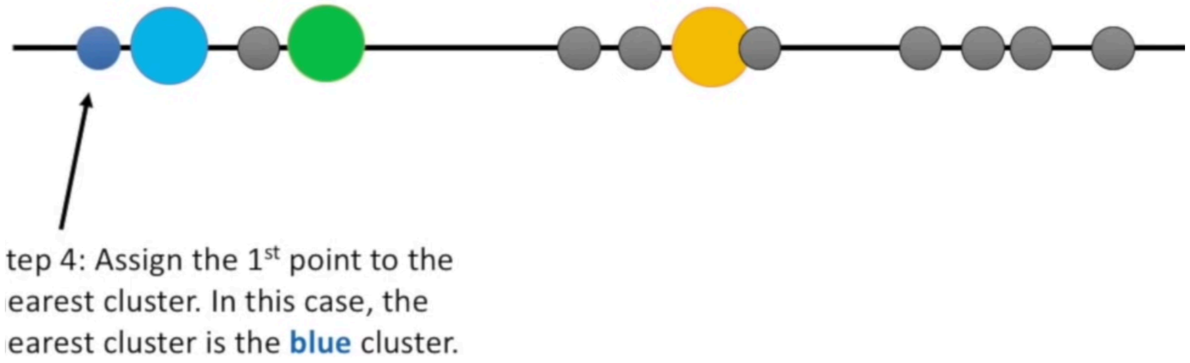
p 3: Measure the distance
tween the 1st point and the three
ial clusters.

now for **yellow**

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



tep 4: Assign the 1st point to the
earest cluster. In this case, the
earest cluster is the **blue** cluster.

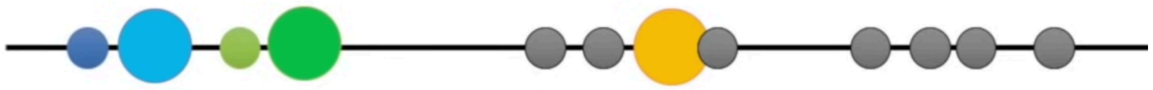
step **4** assign color based on **proximity**

the one with shortest distance is the cluster the grey point should be assigned to…

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**

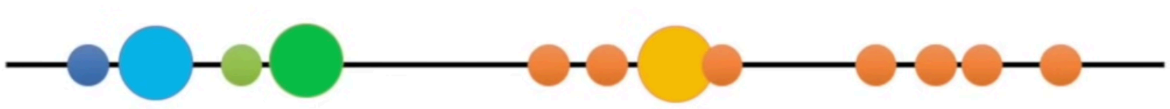


now for second **grey** point

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**



The rest of these points are closest to the orange cluster, so they'll go in that one, too.

... for all the points in the line

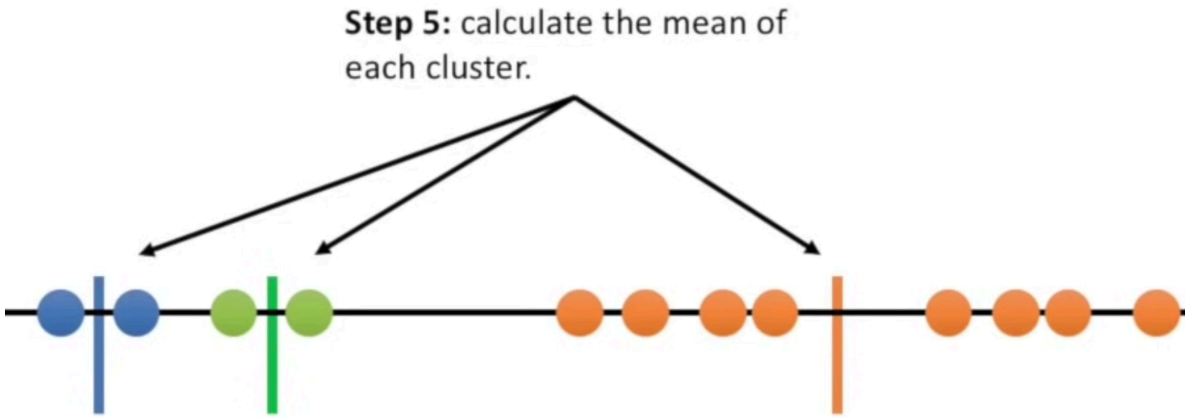
results:

- **blue cluster**: 2 obs
- **green cluster:** 2 obs
- **yellow cluster:** 8 obs

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



Step 5: calculate the mean of each cluster.

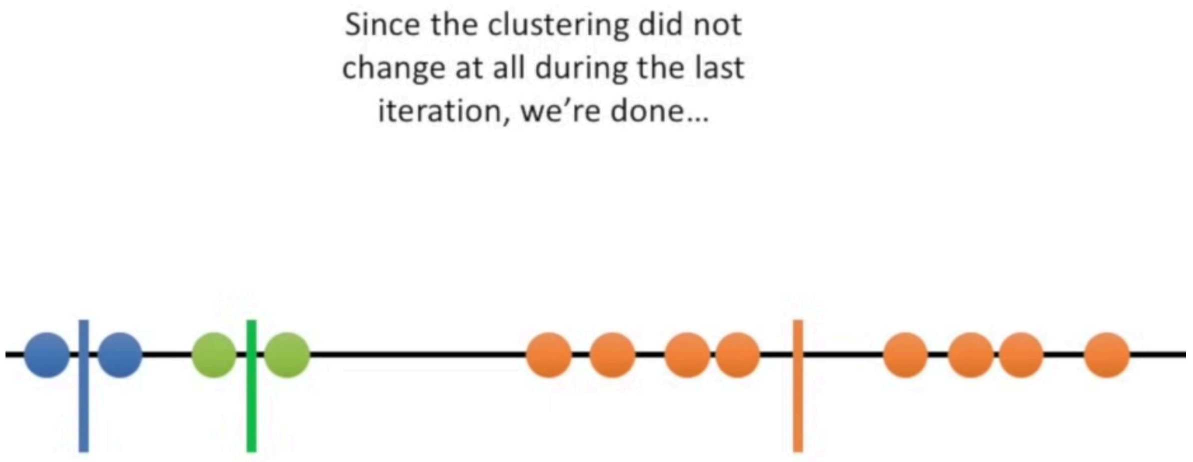
step **5**

compute the mean for each cluster. That's why ***k-means***

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**

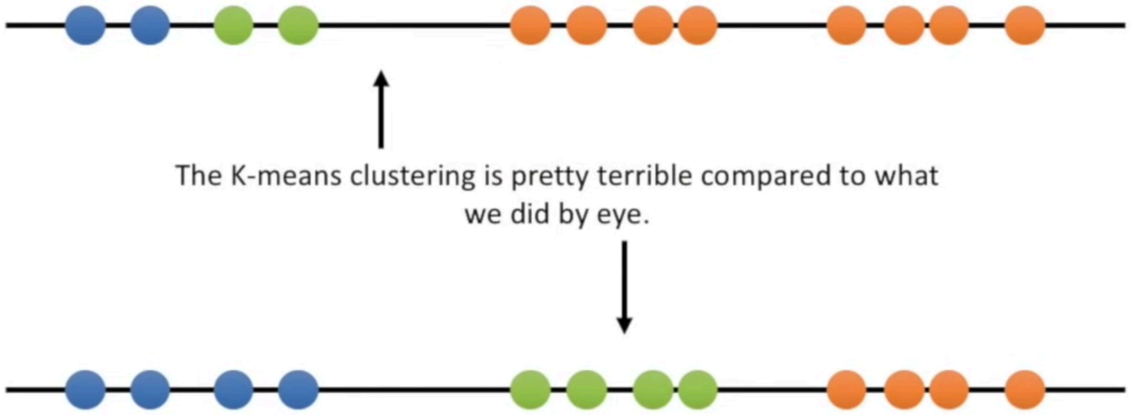


Since the clustering did not change at all during the last iteration, we're done...

then reassingn cluster **based on mean**

©2022 Proprietary and Confidential. All Rights Reserved.

cluster with **heatmaps**



The K-means clustering is pretty terrible compared to what we did by eye.

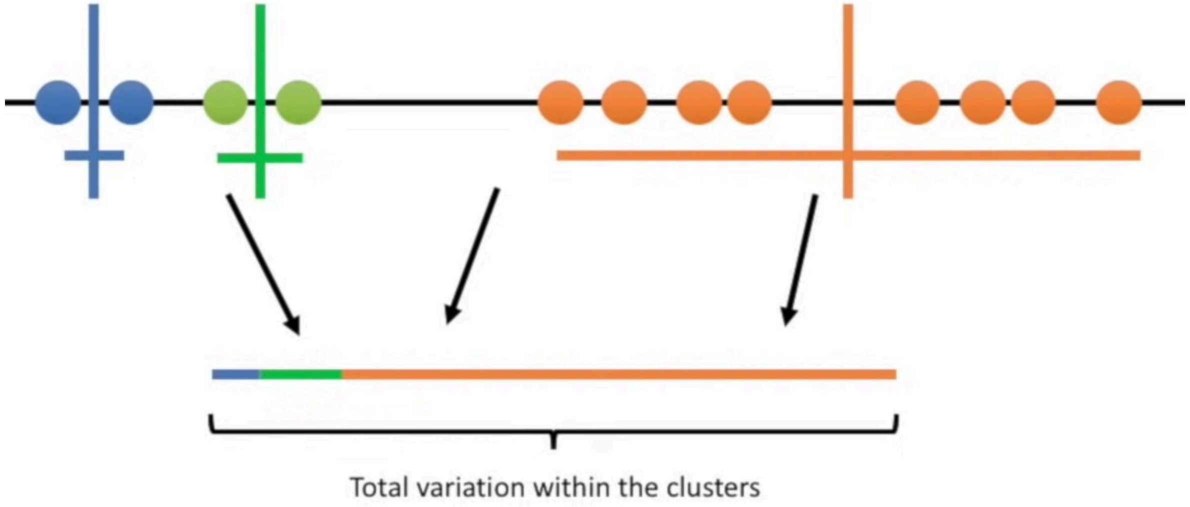
human vs **computer**

we would have done better....

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



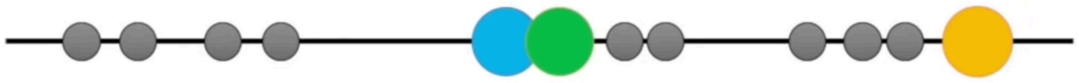
variation within

let's also compute variation within each cluster

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



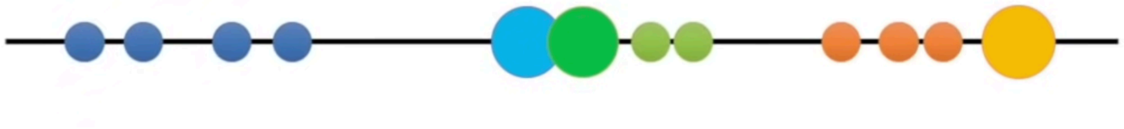
... step 1

The algo restarts...

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**



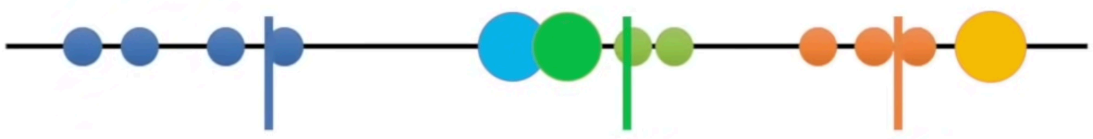
reassign cluster based on **new init**

in this case:
- **blue cluster**: 5 obs
- **green cluster**: 3 obs
- **yellow cluster**: 4 obs

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**

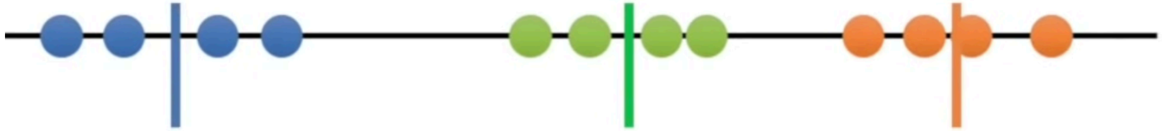


recompute **means..**

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



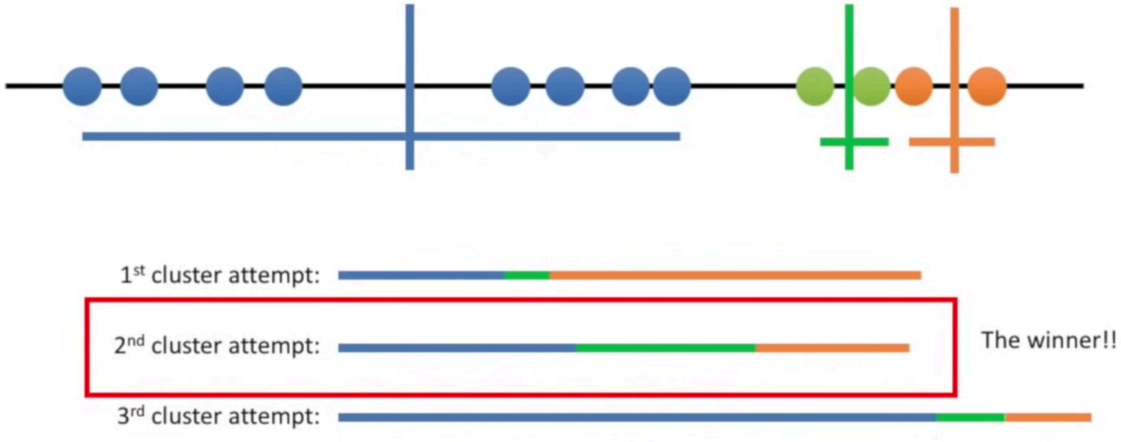
reassign cluster based on **mean**

… well this time is better: *human and computer did the same*

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**

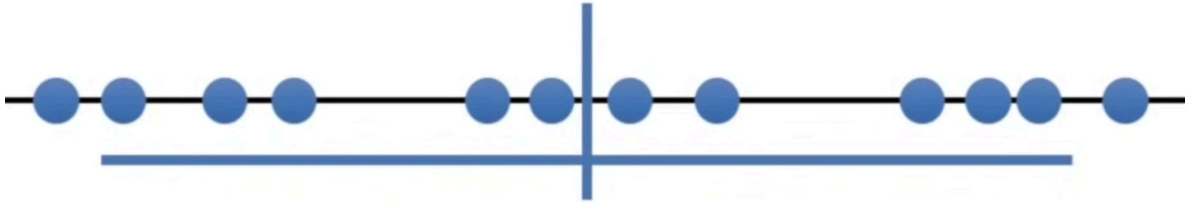


stop when assign = clust

measure differences over attempts (algo iterations)

©2022 Proprietary and Confidential. All Rights Reserved.

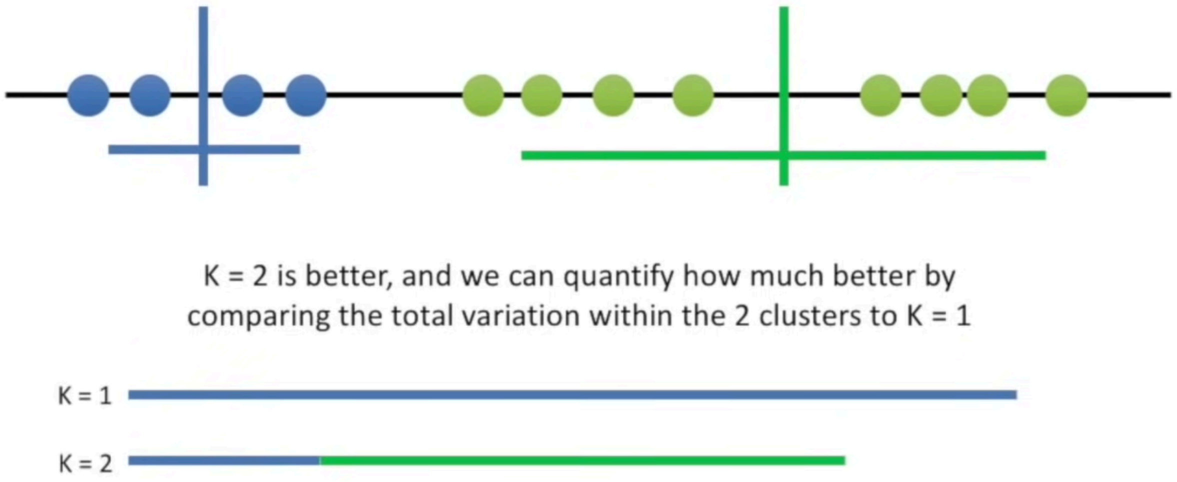
cluster with **heatmaps**



if we would have chosen # k=1

MADE WITH beautiful.ai

cluster with **heatmaps**



K = 2 is better, and we can quantify how much better by comparing the total variation within the 2 clusters to K = 1

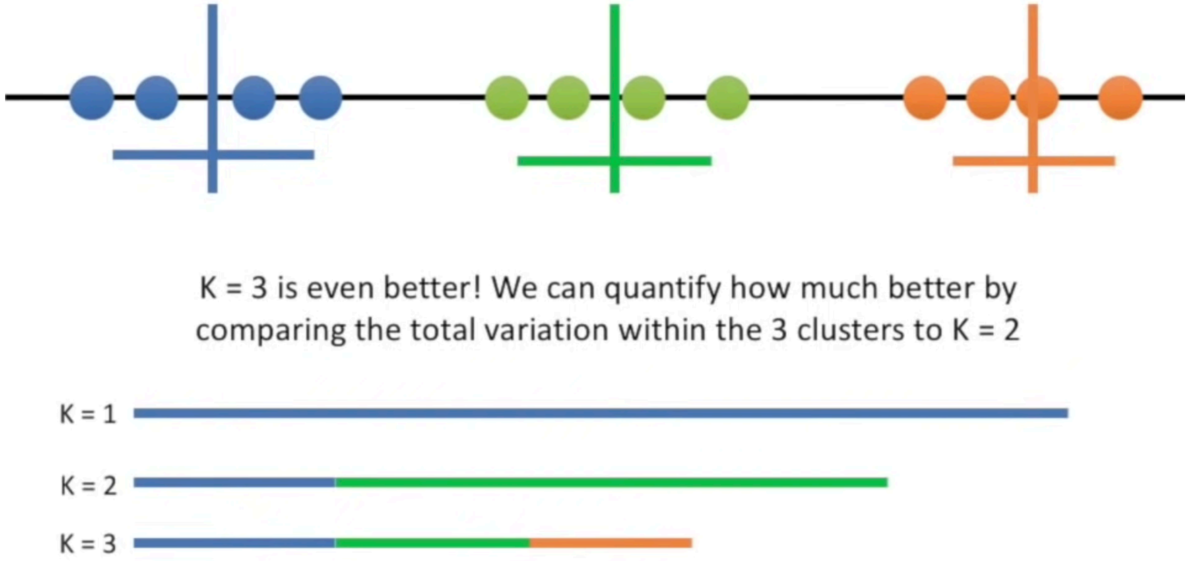
if we would have chosen # **k = 2**

compare below variation within clusters based on number of clusters.

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



K = 3 is even better! We can quantify how much better by comparing the total variation within the 3 clusters to K = 2

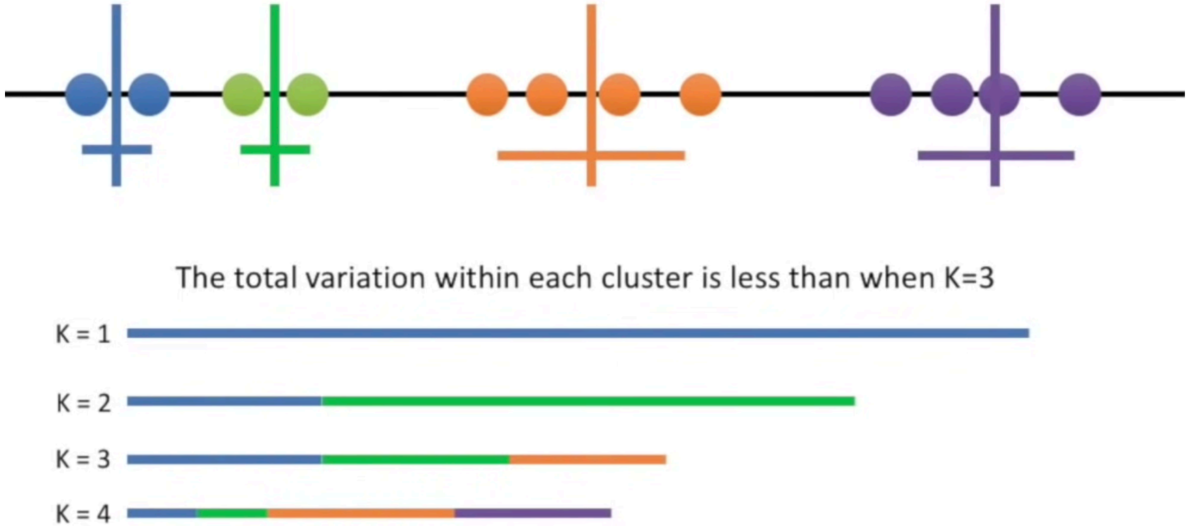
if we would have chosen # **k = 3**

variation withion when k = 3 is actually lower.

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**


The total variation within each cluster is less than when K=3
K = 1
K = 2
K = 3
K = 4

if we would have chosen # **k = 4**

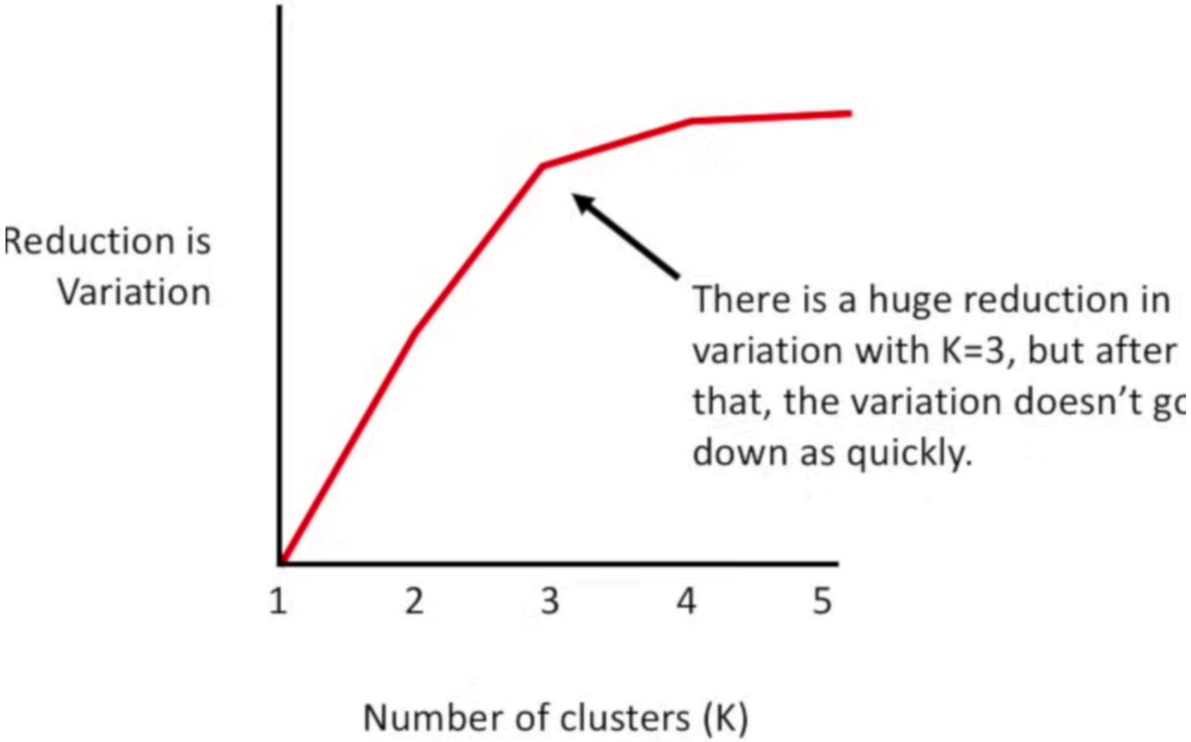
- keeps decreasing.
- that really resembles R2 behaviour, the more params you insert in the model, the better R2
- extreme case 1 clust per obs

we need to find a way to decide which is the best # k.

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



Elbow **method**

popular ML method, plot delta var over # algo iterations (in this case # k).

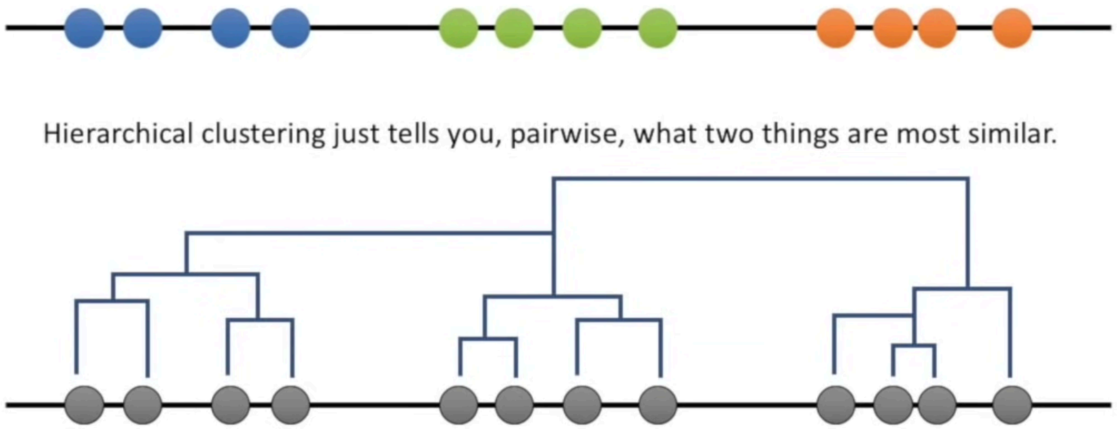
at some point the reduction in variation considerably stop increasing.

the question you should be asking: where should I stop?

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**



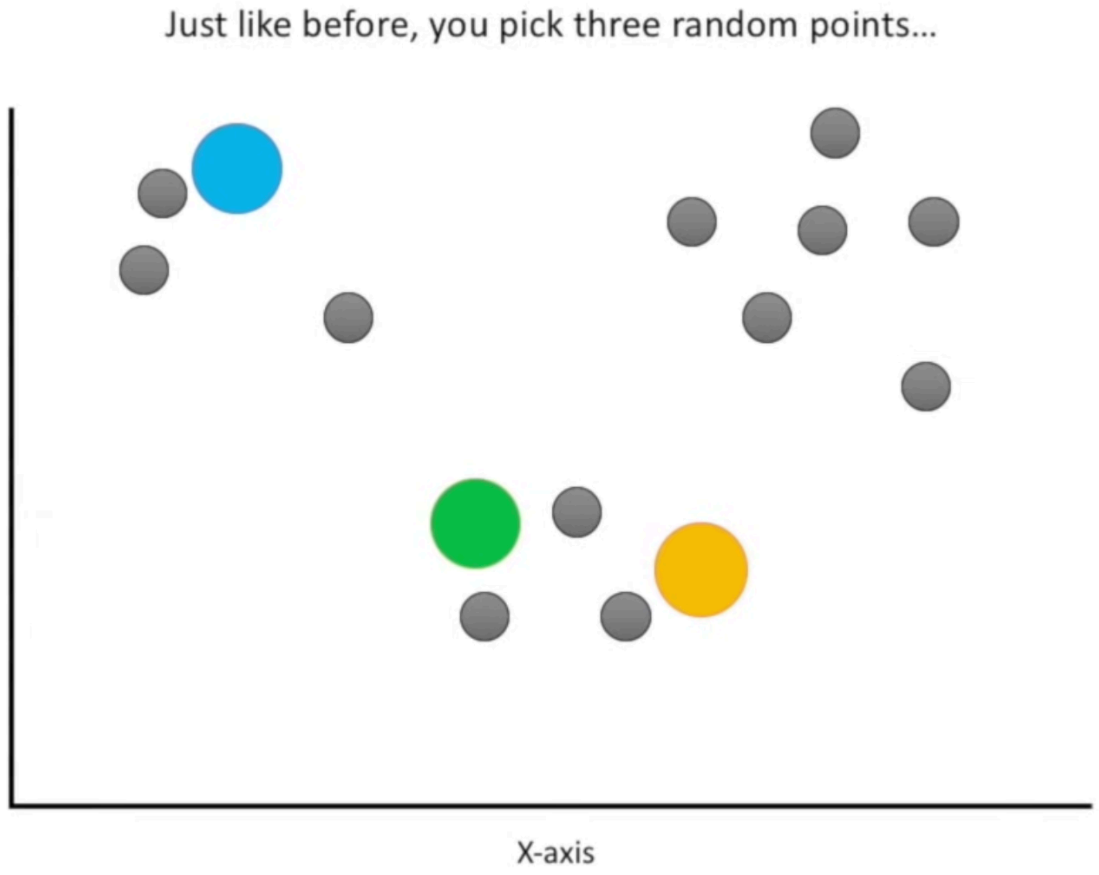
Hierarchical clustering just tells you, pairwise, what two things are most similar.

hclust **vs** k-means

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai

cluster with **heatmaps**

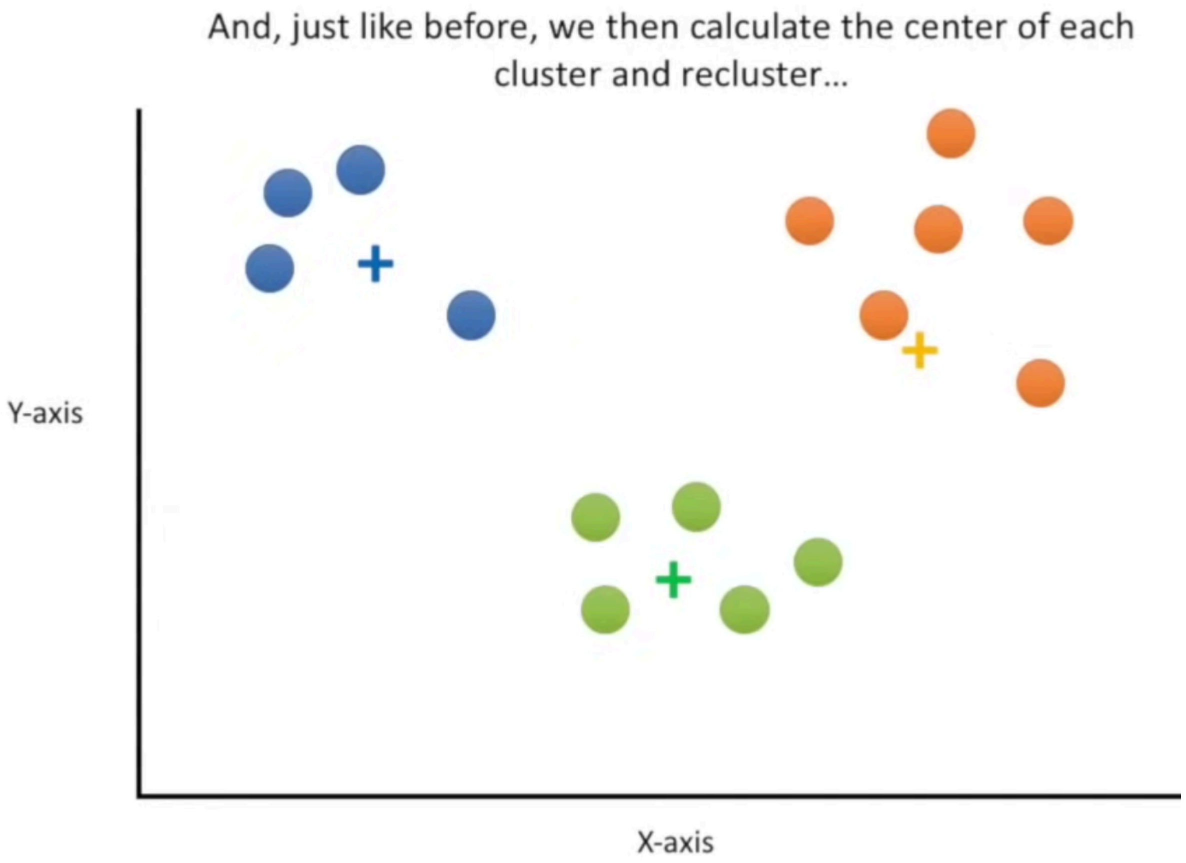


from 2D to 3D

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai

cluster with **heatmaps**



...exactly the same

step 1: randomly assign and observation to each of the 3 clusters.
step 2: compute diff from first assigned to each other point
step 3: assign point to cluster
step 4: iterate over all the clusters
step 5: compute means
step 6: reassign cluster based on mean.

©2022 Proprietary and Confidential. All Rights Reserved.

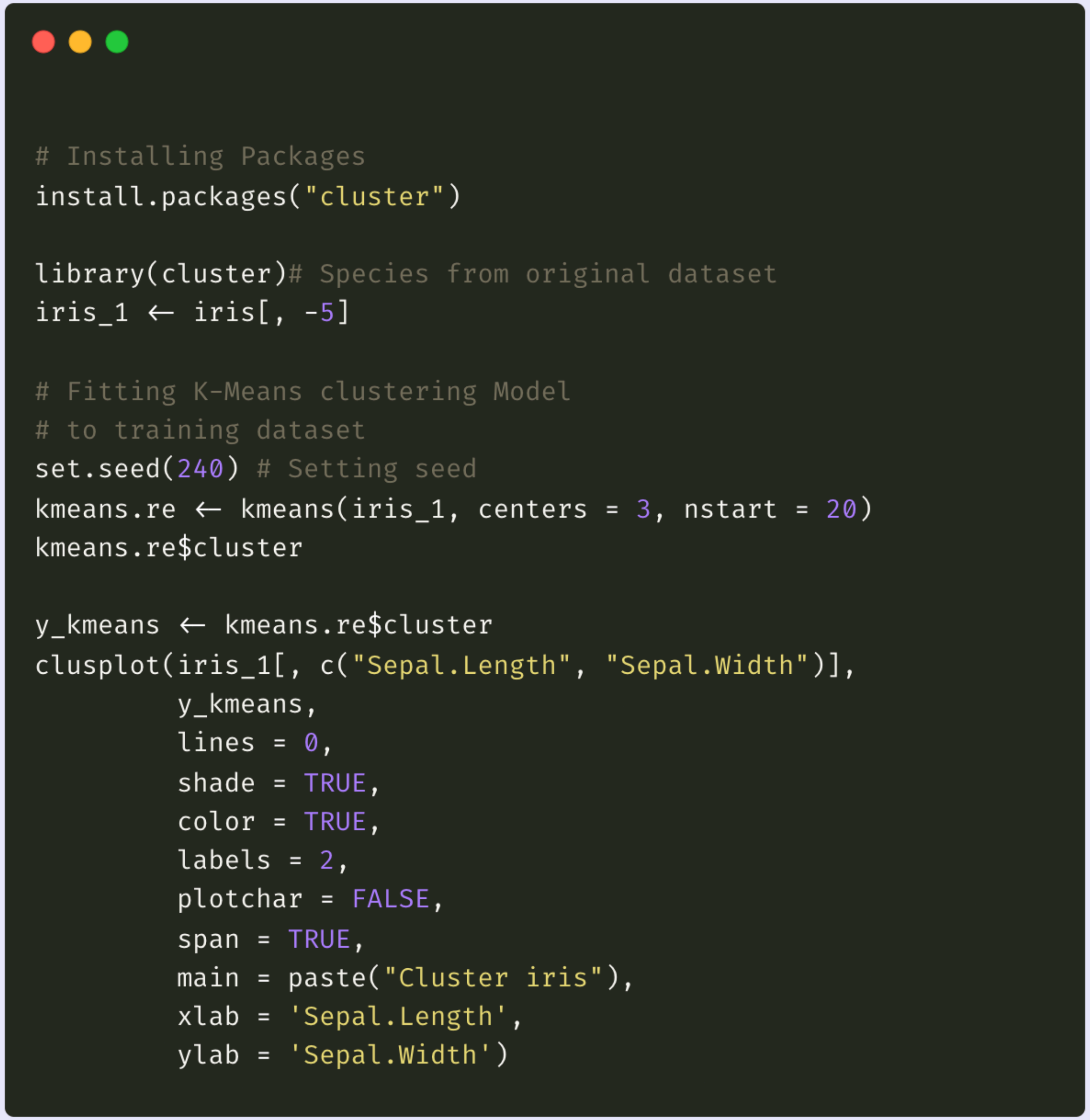
MADE WITH beautiful.ai

Section 3

K-means **R code**

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH beautiful.ai



```r
# Installing Packages
install.packages("cluster")

library(cluster)# Species from original dataset
iris_1 ← iris[, -5]

# Fitting K-Means clustering Model
# to training dataset
set.seed(240) # Setting seed
kmeans.re ← kmeans(iris_1, centers = 3, nstart = 20)
kmeans.re$cluster

y_kmeans ← kmeans.re$cluster
clusplot(iris_1[, c("Sepal.Length", "Sepal.Width")],
         y_kmeans,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 2,
         plotchar = FALSE,
         span = TRUE,
         main = paste("Cluster iris"),
         xlab = 'Sepal.Length',
         ylab = 'Sepal.Width')
```

MADE WITH
beautiful.ai

Section 4

Live coding session!

JUMP TO RSTUDIO!

UNIVERSITÀ
CATTOLICA
del Sacro Cuore

©2022 Proprietary and Confidential. All Rights Reserved.

MADE WITH
beautiful.ai