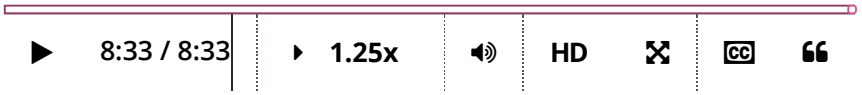


**Video**
Download video file

**Transcripts**
Download SubRip (.srt) file
Download Text (.txt) file

## An Advanced Approach to Finding Cluster Centroids

In this video, we explain how you can find the cluster centroids by using the function "tapply" for each variable in the dataset. While this approach works and is familiar to us, it can be a little tedious when there are a lot of variables. An alternative approach is to use the colMeans function. With this approach, you only have one command for each cluster instead of one command for each variable. If you run the following command in your R console, you can get all of the column (variable) means for cluster 1:

```
colMeans(subset(movies[2:20], clusterGroups == 1))
```

You can repeat this for each cluster by changing the clusterGroups number. However, if you also have a lot of clusters, this approach is not that much more efficient than just using the tapply function.

A more advanced approach uses the "split" and "lapply" functions. The following command will split the data into subsets based on the clusters:

spl = split(movies[2:20], clusterGroups)

Then you can use spl to access the different clusters, because

spl[[1]]

is the same as

subset(movies[2:20], clusterGroups == 1)

so colMeans(spl[[1]]) will output the centroid of cluster 1. But an even easier approach uses the lapply function. The following command will output the cluster centroids for all clusters:

lapply(spl, colMeans)

The lapply function runs the second argument (colMeans) on each element of the first argument (each cluster subset in spl). So instead of using 19 tapply commands, or 10 colMeans commands, we can output our centroids with just two commands: one to define spl, and then the lapply command.

Note that if you have a variable called "split" in your current R session, you will need to remove it with rm(split) so that you can use the split function.

---

In this video, we use the spreadsheet ClusterMeans.ods. This file can be opened in LibreOffice or OpenOffice.

---

## Discussion

Hide Discussion

**Topic:** Unit 6 / Unit 6, Lecture 1, Video 7: Hierarchical Clustering in R

**Add a Post**