```r
# Week 6 - Introduction to Clustering

# Video 6

# After following the steps in the video, load the data into R
movies = read.table("movieLens.txt", header=FALSE,
 sep="|",quote="\"")

str(movies)

# Add column names
colnames(movies) = c("ID", "Title", "ReleaseDate",
 "VideoReleaseDate", "IMDB", "Unknown", "Action", "Adventure",
 "Animation", "Childrens", "Comedy", "Crime", "Documentary", "Drama",
 "Fantasy", "FilmNoir", "Horror", "Musical", "Mystery", "Romance",
 "SciFi", "Thriller", "War", "Western")

str(movies)

# Remove unnecessary variables
movies$ID = NULL
movies$ReleaseDate = NULL
movies$VideoReleaseDate = NULL
movies$IMDB = NULL

# Remove duplicates
movies = unique(movies)

# Take a look at our data again:
str(movies)



# Video 7

# Compute distances
distances = dist(movies[2:20], method = "euclidean")

# Hierarchical clustering
clusterMovies = hclust(distances, method = "ward")

# Plot the dendrogram
plot(clusterMovies)

# Assign points to clusters
clusterGroups = cutree(clusterMovies, k = 10)

#Now let's figure out what the clusters are like.
```

```r
# Let's use the tapply function to compute the percentage of movies
 in each genre and cluster

tapply(movies$Action, clusterGroups, mean)
tapply(movies$Romance, clusterGroups, mean)

# We can repeat this for each genre. If you do, you get the results
 in ClusterMeans.ods


# Find which cluster Men in Black is in.

subset(movies, Title=="Men in Black (1997)")
clusterGroups[257]

# Create a new data set with just the movies from cluster 2
cluster2 = subset(movies, clusterGroups==2)

# Look at the first 10 titles in this cluster:
cluster2$Title[1:10]
```