**3. Clustering. Use the “movieLens.txt” data in Lecture 3 and perform cluster analysis of the 1664 non-duplicate movies.**

*movies = read.table("movieLens.txt", header=FALSE, sep="|", quote="\"")*

*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")*

*movies$ID = NULL*

*movies$ReleaseDate = NULL*

*movies$VideoReleaseDate = NULL*

*movies$IMDB = NULL*

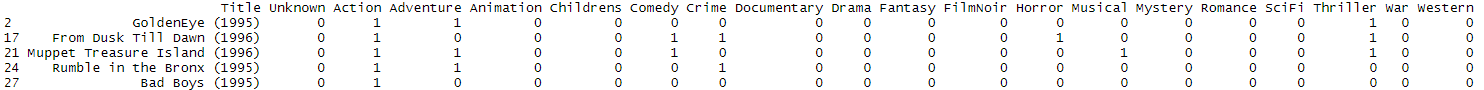
*movies = unique(movies)*

**a. Apply k-means clustering to the 19 binary (0/1) variables that capture movies attributes. To ensure the analysis results are stable, set the rounds of re-start to 100 and the number of maximum iterations in each round to 200. Set k=6 and show 5 movies in each cluster.**

*Movie.seg = kmeans(scale(movies[,2:20]), centers=6,nstart = 100,iter.max=200)*

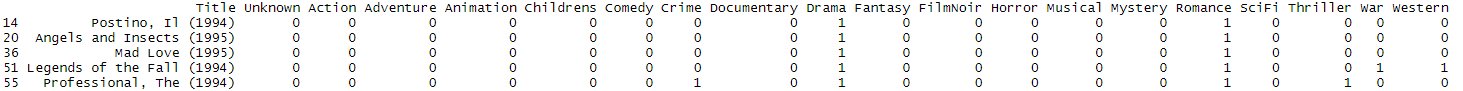
*cluster1 = subset(movies, Movie.seg$cluster==1)*

*head(cluster1,n=5)*

**

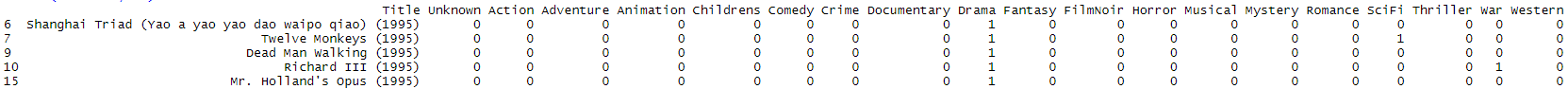
*cluster2 = subset(movies, Movie.seg$cluster==2)*

*head(cluster2,n=5)*

**

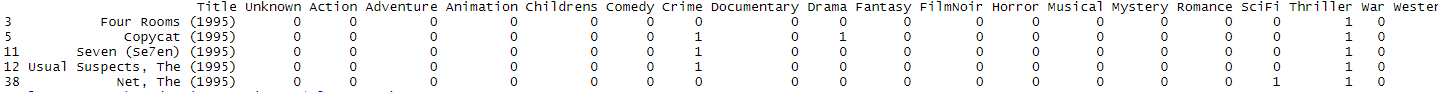
*cluster3 = subset(movies, Movie.seg$cluster==3)*

*head(cluster3,n=5)*

**

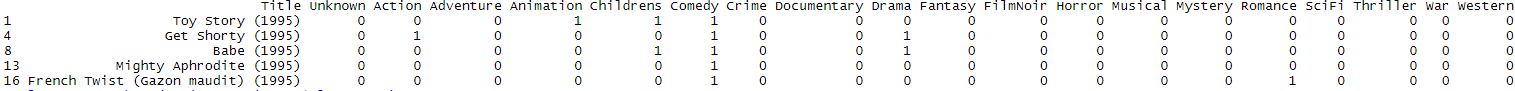
*cluster4 = subset(movies, Movie.seg$cluster==4)*

*head(cluster4,n=5)*

**

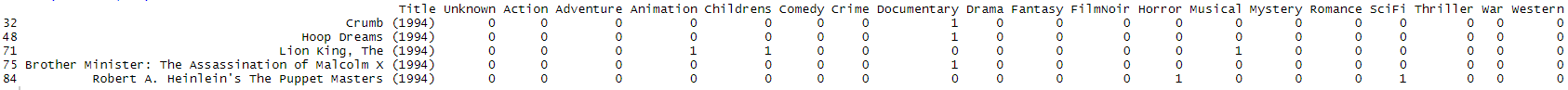
*cluster5 = subset(movies, Movie.seg$cluster==5)*

*head(cluster5,n=5)*

**

*cluster6 = subset(movies, Movie.seg$cluster==6)*

*head(cluster6,n=5)*

**

由上圖可以分別看出每一群的前五筆電影，分別為

第1群：GoldenEye (1995)、From Dusk Till Dawn (1996)、Muppet Treasure Island (1996)、Rumble in the Bronx (1995)、Bad Boys (1995)

第2群：Postino, Il (1994)、Angels and Insects (1995)、Mad Love (1995)、Legends of the Fall (1994)、Professional, The (1994)

第3群：Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)、Twelve Monkeys (1995)、Dead Man Walking (1995)、Richard III (1995)、Mr. Holland's Opus (1995)

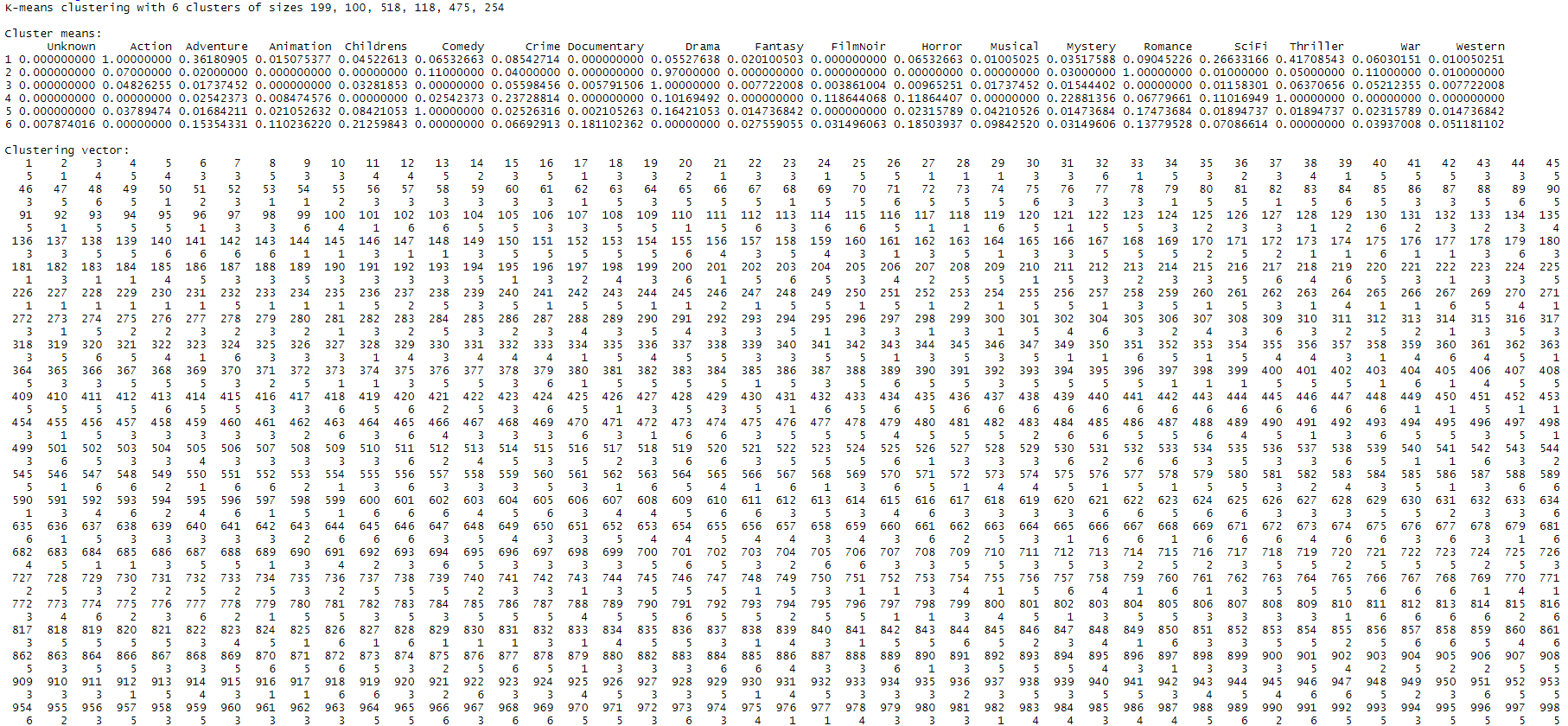
第4群：Four Rooms (1995)、Copycat (1995)、Seven (Se7en) (1995)、Usual Suspects, The (1995)、Net, The (1995)

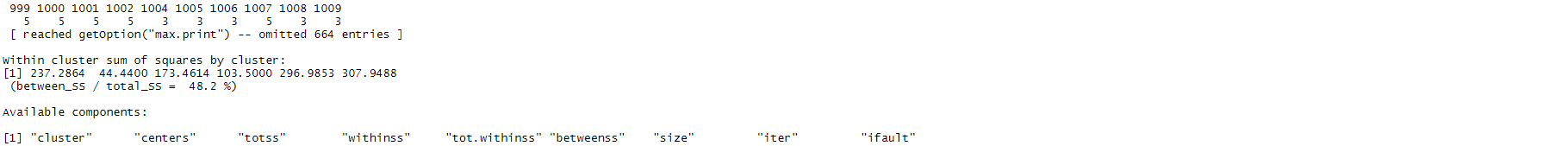
第5群：Toy Story (1995)、Get Shorty (1995)、Babe (1995)、Mighty Aphrodite (1995)、French Twist (Gazon maudit) (1995)

第6群：Crumb (1994)、Hoop Dreams (1994)、Lion King, The (1994)、Brother Minister: The Assassination of Malcolm X (1994)、Robert A. Heinlein's The Puppet Masters (1994)

**b. In class we identified a couple of clusters as Action/Romance (see Lecture 3 R code). Try to make sense and label each of the 6 clusters in part a.**

*Movie.seg*





由圖可以將第1群歸類為Action；第2群歸類為Romance；第3群歸類為Drama；第4群歸類為Thriller；第5群歸類為Comedy；而第6群因為平均值都過低，因此難以將其歸類為任何一類。

**c. Are you happy with k=6 and the categorization results in part b? Would you increase or decrease k? Explain why you increase or decrease k and articulate the changes in movie clusters. What is the impact of those changes for making recommendations?**

*k=6*

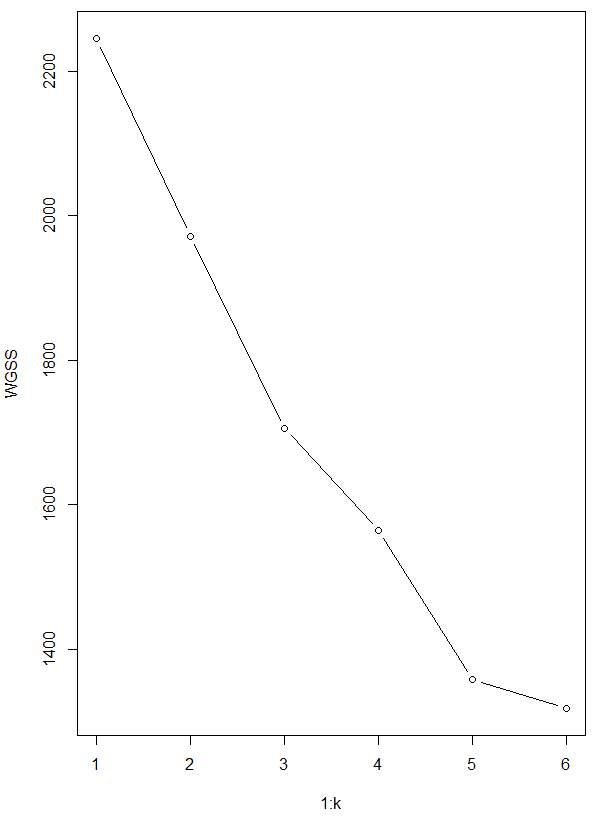
*WGSS=c()*

*for(i in 1:k){*

*WGSS[i]=sum(kmeans(movies[,2:20], centers=i)$withinss)*

*}*

*plot(1:k, WGSS, type="b")*

**

由圖可以看出我們大約只需選擇5群即可。