Q1. Cosmetics Purchases

Trans. #	Bag	Blush	Nail Polish	Brushes	Concealer	Eyebrow Pencils	Bronzer
1	0	1	1	1	1	0	1
2	0	0	1	0	1	0	1
3	0	1	0	0	1	1	1
4	0	0	1	1	1	0	1
5	0	1	0	0	1	0	1
6	0	0	0	0	1	0	0
7	0	1	1	1	1	0	1
8	0	0	1	1	0	0	1
9	0	0	0	0	1	0	0
10	1	1	1	1	0	0	0
11	0	0	1	0	0	0	1
12	0	0	1	1	1	0	1

a.The matrix form shows that each row represents a transaction, and the columns are different cosmetic products. The values of the matrix form show whether that kind of the cosmetics is in the transaction or not. For example: In transaction 6, the customer bought the concealer, so the value is 1. Because the customer didn't buy like bag, brushes, the values would be 0

TABLE 14.15 ASSOCIATION RULES FOR COSMETICS PURCHASES DATA

	lhs	rhs	support	confidence	lift
1	{Blush, Concealer, Mascara, Eye.shadow, Lipstick}	=> {Eyebrow.Pencils}	0.013	0.3023255814	7.198228128
2	{Trans., Blush, Concealer, Mascara, Eye.shadow, Lipstick}	=> {Eyebrow.Pencils}	0.013	0.3023255814	7.198228128
3	{Blush, Concealer, Mascara, Lipstick}	=> {Eyebrow.Pencils}	0.013	0.2888888889	6.878306878
4	{Trans., Blush, Concealer, Mascara, Lipstick}	=> {Eyebrow.Pencils}	0.013	0.2888888889	6.878306878
5	{Blush, Concealer, Eye.shadow, Lipstick}	=> {Eyebrow.Pencils}	0.013	0.2826086957	6.728778468
6	{Trans., Blush, Concealer, Eye.shadow, Lipstick}	=> {Eyebrow.Pencils}	0.013	0.2826086957	6.728778468

b.

I.The confidence of the first row means that given we see blush, concealer, mascara, eye shadow and lipstick in a transaction, what's the probability that we see eyebrow pencil? The calculation would be :

$$Confidence = P(eyebrow\ pencil \mid blush, concealer, mascara, eyeshadow, lipstick)$$

$$= \frac{(number\ of\ transactions\ with\ eyebrow\ pencil, blush, concealer, mascara, eye\ shadow, lipstick)}{(number\ of\ transactions\ with\ blush, concealer, mascara, eye\ shadow, lipstick)}$$

$$= 0.3023$$

II. Support ensures that items (blush, concealer, mascara, eye shadow, lipstick) that occur relatively frequently in transactions.The calculation would be :

$$Support = P(blush, concealer, mascara, eyeshadow, lipstick)$$

$$= \frac{(number\ of\ transactions\ with\ blush, concealer, mascara, eye\ shadow, lipstick)}{(total\ number\ of\ transactions)}$$

$$= 0.0130$$

III. Lift examines how much the rule improves occurrence of the consequent item without the rule. That is, the probability buying consequent increases x percent when we see antecedent.The calculation would be :

$$Lift = \frac{P(eyebrow\ pencil \mid blush, concealer, mascara, eye\ shadow, lipstick)}{P(eyebrow\ pencil)}$$

$$= \frac{P(eyebrow\ pencil, blush, concealer, mascara, eye\ shadow, lipstick)}{P(eyebrow\ pencil))} \times P(blush, concealer, mascara, eye\ shadow, lipstick)$$

$$= 7.1982$$

IV. If blush, concealer, mascara, eye shadow and lipstick are purchased, then we are 30.23% confident that eyebrow pencil will also be purchased. This rule would be 719% better than purchasing eyebrow pencil only.

c.

i. The first rule of the output shows that if brushes are purchased, then we assure that the nail polish will also be purchased. Also, It would be 357% better than purchasing nail polish only.

Look at the second and third rule, both the antecedents include blush and eye shadow, and the consequent is also the same(mascara). These two rules both show over 90% confidence. The only difference between these two is that the antecedent of the second rule adds concealer.

ii.

```
     lhs                              rhs           support confidence lift     count
[1]  {Brushes}                     => {Nail.Polish} 0.149   1.0000000  3.571429 149
[2]  {Blush,Concealer,Eye.shadow}  => {Mascara}     0.119   0.9596774  2.688172 119
[3]  {Blush,Eye.shadow}            => {Mascara}     0.169   0.9285714  2.601040 169
```

[2],[3] 只差[3]多了Concealer

```
[4]  {Nail.Polish,Eye.shadow}      => {Mascara}     0.119   0.9083969  2.544529 119
[5]  {Concealer,Eye.shadow}        => {Mascara}     0.179   0.8905473  2.494530 179
```

[2],[5] 重複只差[2]多了Blush

```
[6]  {Bronzer,Eye.shadow}          => {Mascara}     0.124   0.8794326  2.463397 124
[12] {Bronzer,Mascara}             => {Eye.shadow}  0.124   0.9051095  2.375615 124
```

[6],[12] 重複只是順序顛倒

```
[7]  {Concealer,Eye.shadow,Eyeliner} => {Mascara}     0.114   0.8769231  2.456367 114
[8]  {Blush,Mascara}                 => {Eye.shadow}  0.169   0.9184783  2.410704 169
[9]  {Eye.shadow,Lipstick}           => {Mascara}     0.110   0.8527132  2.388552 110
[10] {Mascara,Lipstick}              => {Eye.shadow}  0.110   0.9090909  2.386065 110
```

[9],[10] 重複只是順序顛倒

```
[11] {Blush,Concealer,Mascara}       => {Eye.shadow}  0.119   0.9083969  2.384244 119
```

刪除多餘的規則後如下:

```
     lhs                                rhs           support confidence lift     count
[1]  {Brushes}                       => {Nail.Polish} 0.149   1.0000000  3.571429 149
[2]  {Blush,Concealer,Eye.shadow}    => {Mascara}     0.119   0.9596774  2.688172 119
[4]  {Nail.Polish,Eye.shadow}        => {Mascara}     0.119   0.9083969  2.544529 119
[6]  {Bronzer,Eye.shadow}            => {Mascara}     0.124   0.8794326  2.463397 124
[7]  {Concealer,Eye.shadow,Eyeliner} => {Mascara}     0.114   0.8769231  2.456367 114
[9]  {Eye.shadow,Lipstick}           => {Mascara}     0.110   0.8527132  2.388552 110
```

這些規則可能是比較重要且沒有重複性的的，當然購物籃分析中，左邊所包含的商品及右邊推薦的商品，其相關性可能也是一件能夠參考的事情，如[2]與[3]規則中，

```
[2]  {Blush,Concealer,Eye.shadow}    => {Mascara}
[3]  {Blush,Eye.shadow}              => {Mascara}
```

[3] 其實也有其重要的涵義，有 {Blush,Eye.shadow} 但沒有 {Concealer,Eye.shadow}，可以說明 {Blush,Eye.shadow} 的關聯性較佳，並且通常買這兩個商品的人也會買 {Mascara}，而[9]跟[10]也有同樣概念，有

```
[9]  {Eye.shadow,Lipstick}           => {Mascara}
[10] {Mascara,Lipstick}              => {Eye.shadow}
```

但卻沒有 {Mascara, Eye.shadow} => {Lipstick}，表示 {Mascara, Eye.shadow} 兩者的關聯性可能較不高。

Q2. Course ratings

TABLE 14.16 RATINGS OF ONLINE STATISTICS COURSES: 4 = BEST, 1 = WORST, BLANK = NOT TAKEN

	SQL	Spatial	PA 1	DM in R	Python	Forecast	R Prog	Hadoop	Regression
L N	4				3	2	4		2
M H	3	4			4				
J H	2	2							
E N	4			4			4		3
D U	4	4							
F L		4							
G L		4							
A H		3							
S A			4						
R W			2					4	
B A			4						
M G			4			4			
A F			4						
K G			3						
D S	4			2			4		

a.
The equation below is the correlation between two users' ratings:

$$Corr(U_1, U_2) = \frac{\sum (r_{1,i} - \bar{r}_1)(r_{2,i} - \bar{r}_2)}{\sqrt{\sum (r_{1,i} - \bar{r}_1)^2}\sqrt{(r_{2,i} - \bar{r}_2)^2}}$$

	L.N.	M.H.	J.H.	E.N.	D.U.	D.S.
E.N.	0.8704	-1.0000	NaN	NaN	NaN	-1.570092E-16

The single nearest student to E.N is L.N （Correlation : 0.8704）

b. Our answer is Python.
L.N is the nearest student to E.N.. Except for the common courses what L.N and E.N. complete, Python is the highest rating course in L.N's completed course.

c. cosine similarity between users

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	0.9600													
3	1.0000	0.9899												
4	0.9891	1.0000	1.0000											
5	1.0000	0.9899	1.0000	1.0000										
6	NA	1.0000	1.0000	NA	1.0000									
7	NA	1.0000	1.0000	NA	1.0000	1.0000								
8	NA	1.0000	1.0000	NA	1.0000	1.0000	1.0000							
9	NA	NA	NA	NA	NA	NA	NA	NA						
10	NA	NA	NA	NA	NA	NA	NA	NA	1.0000					
11	NA	NA	NA	NA	NA	NA	NA	NA	1.0000	1.0000				
12	1.0000	NA	NA	NA	NA	NA	NA	NA	1.0000	1.0000	1.0000			
13	NA	NA	NA	NA	NA	NA	NA	NA	1.0000	1.0000	1.0000	1.0000		
14	NA	NA	NA	NA	NA	NA	NA	NA	1.0000	1.0000	1.0000	1.0000	1.0000	
15	1.0000	1.0000	1.0000	0.9623	1.0000	NA	NA	NA	NA	NA	NA	NA	NA	NA

d.
Based on the cosine similarities, the nearest students to E.N is M.H. , J.H. and D.U. When we look up the common courses what they complete, Spatial is the only course which M.H. , J.H. and D.U. complete but E.N does not. That is the reason Spatial course should be recommended to E.N.

e.
兩者的差距在於，correlation有扣除平均值，而cosine則是直接拿原始的值來算相似度，而扣除平均值是為了把每個人評分的嚴謹程度算進去，像是對同一門課評價都為5分的人，其中可能包含:
1.給分嚴謹的人，他很少給出5分，這個5分的特殊性就較高。
2.給分彈性的人，他時常給出5分，這個5分的特殊性就較低。
透過扣掉平均，可以把這樣的狀況給算進去(第1種人扣掉平均有較高的數字)，
而cosine則是不去扣掉平均，可以讓運算更加簡單，只看向量的關係性。

f.
i.
EN有修的課是SQL/DM.in.R/R.Prog/Regression
計算SQL/DM.in.R/R.Prog/Regression與其他課的相似度，並去預測EN對於其他課的喜好度，並且要有同時修兩堂課，都有評分的資料。
SQL + *Spatial *Python *Forecast
R.Prog + *Python *Forecast
Regression + *Python *Forecast
並且要找出有至少兩組評分的有兩組 (SQL, Spatial) (SQL, Python)並去計算其相似度。
ii.
推薦python由於EN有修的課，同樣修這些課的人，也有同樣修python並且評價不錯(3分/4分相比Forecast2/4分)
計算其相似度，透過相似度，推薦Spatial。

	row	column	cor
1	SQL	Spatial	0.8181
7	SQL	Python	-1.000

g.
透過item_based做預測最後得到的結果為下圖

	SQL	Spatial	PA.1	DM.in.R	Python	Forecast	R.Prog	Hadoop	Regression
[1,]	NA	4	2	3.26231	NA	NA	NA	NA	NA
[2,]	NA	NA	NA	3.00000	NA	3.000000	3	NA	3.000000
[3,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[4,]	NA	4	NA	NA	4	3.666667	NA	NA	NA
[5,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[6,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[7,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[8,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[9,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[10,]	NA	NA	NA	NA	NA	2.000000	NA	NA	NA
[11,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[12,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[13,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[14,]	NA	NA	NA	NA	NA	NA	NA	NA	NA
[15,]	NA	4	NA	NA	4	4.000000	NA	NA	3.296585

可以推薦 Spatial , Python , Forecast三門課
但是判斷R的算法可能與老師教得有些不同，由於Forcast無法計算相關性但卻出現在推薦並且根據上題 Spatial 及 Python 的相關性，推薦較大相關性的 Spatial 給E.N.

Q3. Clustering

a.

- head(cluster1,n=5)

	Title	Unknown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thriller	War	Western
2	GoldenEye (1995)	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
17	From Dusk Till Dawn (1996)	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1	0	0
21	Muppet Treasure Island (1996)	0	1	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
24	Rumble in the Bronx (1995)	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
27	Bad Boys (1995)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- head(cluster2,n=5)

	Title	Unknown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thriller	War	Western
14	Postino, Il (1994)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
20	Angels and Insects (1995)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
36	Mad Love (1995)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
51	Legends of the Fall (1994)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1
55	Professional, The (1994)	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0

- head(cluster3,n=5)

	Title	Unknown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thriller	War	Western
6	Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	Twelve Monkeys (1995)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
9	Dead Man Walking (1995)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	Richard III (1995)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
15	Mr. Holland's Opus (1995)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

- head(cluster4,n=5)

	Title	Unknown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thriller	War	Wester
3	Four Rooms (1995)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
5	Copycat (1995)	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	
11	Seven (Se7en) (1995)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	
12	Usual Suspects, The (1995)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	
38	Net, The (1995)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	

- head(cluster5,n=5)

	Title	Unknown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thriller	War	Western
1	Toy Story (1995)	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Get Shorty (1995)	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
8	Babe (1995)	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
13	Mighty Aphrodite (1995)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
16	French Twist (Gazon maudit) (1995)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0

- head(cluster6,n=5)

	Title	Unknown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thriller	War	Western
32	Crumb (1994)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
48	Hoop Dreams (1994)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
71	Lion King, The (1994)	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
75	Brother Minister: The Assassination of Malcolm X (1994)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
84	Robert A. Heinlein's The Puppet Masters (1994)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0

According to the chart above we can know that the first five movies in every cluster.

cluster1：GoldenEye (1995)、From Dusk Till Dawn (1996)、Muppet Treasure Island (1996)、Rumble in the Bronx (1995)、Bad Boys (1995)

cluster2：Postino, Il (1994)、Angels and Insects (1995)、Mad Love (1995)、Legends of the Fall (1994)、Professional, The (1994)

cluster3：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)

cluster4：Four Rooms (1995)、Copycat (1995)、Seven (Se7en) (1995)、Usual Suspects, The (1995)、Net, The (1995)

cluster5：Toy Story (1995)、Get Shorty (1995)、Babe (1995)、Mighty Aphrodite (1995)、French Twist (Gazon maudit) (1995)

cluster6：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.

According to the chart above we can know that：

Cluster 1 can be classified as an action ;

Cluster 2 can be classified as an Romance ;

Cluster 3 can be classified as an Drama ;

Cluster 4 can be classified as an Thriller ;

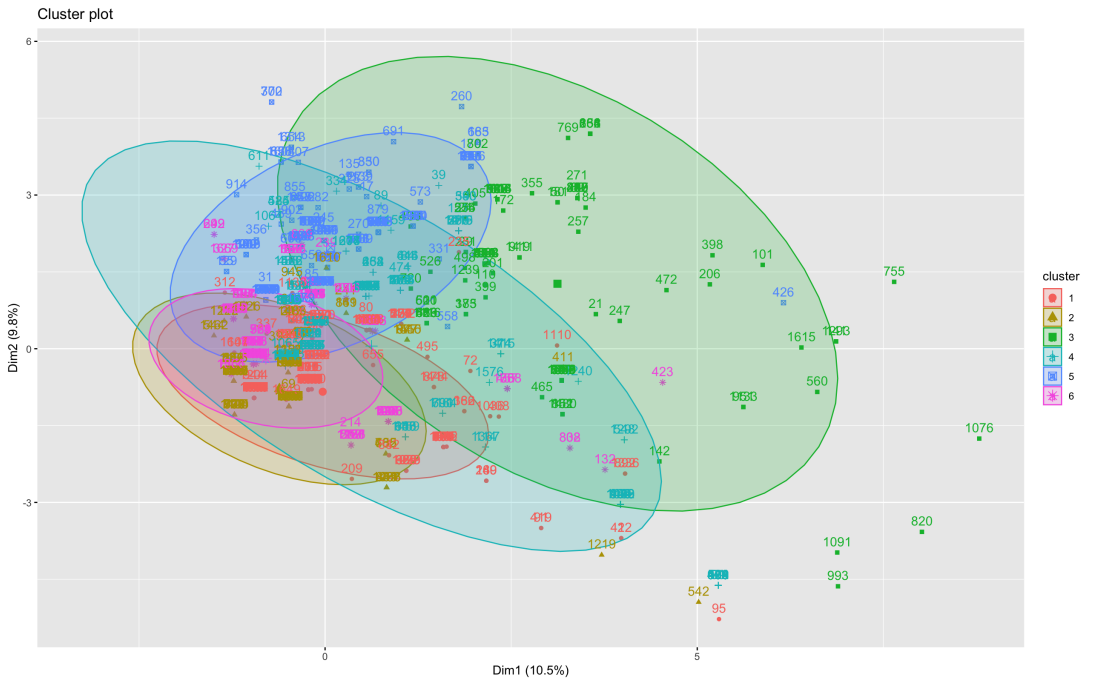
Cluster 5 can be classified as an Comedy and, especially,

the average of Cluster 6 is too low to be classified into any category.

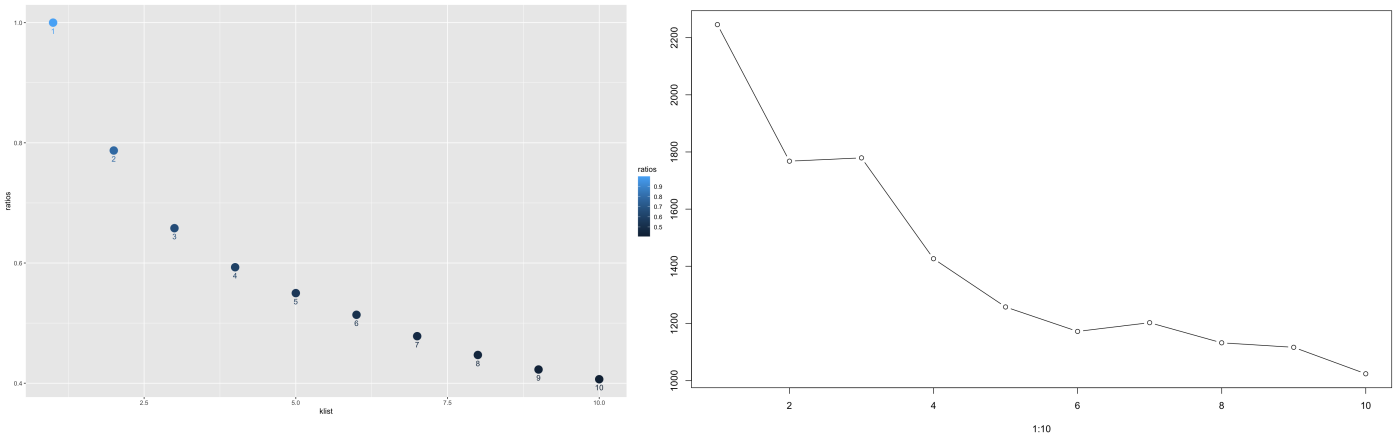
```
K-means clustering with 6 clusters of sizes 199, 100, 518, 118, 475, 254

Cluster means:
      Unknown      Action  Adventure   Animation   Childrens      Comedy       Crime Documentary       Drama     Fantasy    FilmNoir      Horror     Musical     Mystery     Romance       SciFi    Thriller         War     Western
1 0.000000000 1.00000000 0.36180905 0.015075377 0.04522613 0.06532663 0.08542714 0.000000000 0.05527638 0.020100503 0.000000000 0.06532663 0.01005025 0.03517588 0.09045226 0.26633166 0.41708543 0.06030151 0.010050251
2 0.000000000 0.07000000 0.02000000 0.000000000 0.00000000 0.11000000 0.04000000 0.000000000 0.97000000 0.000000000 0.000000000 0.00000000 0.00000000 0.03000000 1.00000000 0.01000000 0.05000000 0.11000000 0.010000000
3 0.000000000 0.04826255 0.01737452 0.000000000 0.03281853 0.00000000 0.05598456 0.005791506 1.00000000 0.007722008 0.003861004 0.00965251 0.01737452 0.01544402 0.00000000 0.01158301 0.06370656 0.05212355 0.007722008
4 0.000000000 0.00000000 0.02542373 0.008474576 0.00000000 0.02542373 0.23728814 0.000000000 0.10169492 0.000000000 0.118644068 0.11864407 0.00000000 0.22881356 0.06779661 0.11016949 1.00000000 0.00000000 0.000000000
5 0.000000000 0.03789474 0.01684211 0.021052632 0.08421053 1.00000000 0.02526316 0.002105263 0.16421053 0.014736842 0.000000000 0.02315789 0.04210526 0.01473684 0.17473684 0.01894737 0.01894737 0.02315789 0.014736842
6 0.007874016 0.00000000 0.15354331 0.110236220 0.21259843 0.00000000 0.06692913 0.181102362 0.00000000 0.027559055 0.031496063 0.18503937 0.09842520 0.03149606 0.13779528 0.07086614 0.00000000 0.03937008 0.051181102

Clustering vector:
   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
   5   1   4   5   4   3   3   5   3   3   4   4   5   2   3   5   1   3   3   2   1   3   3   1   5   5   1   1   1   3   3   6   1   5   3   2   3   4   1   5   5   5   3   3   5
  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
   3   5   6   5   1   2   3   1   1   2   3   3   3   3   3   3   1   5   3   5   5   5   1   5   5   6   5   5   5   6   3   3   3   1   5   5   1   5   6   5   3   3   5   6   5
  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135
   5   1   5   5   1   3   6   4   1   6   6   5   5   3   6   1   3   6   6   1   5   6   3   6   6   5   1   1   5   3   2   3   3   1   2   6   2   3   2   3   4
 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
   3   3   5   6   6   6   6   1   1   3   1   1   3   5   5   5   5   6   4   3   5   4   3   1   3   5   1   3   3   5   5   2   5   2   1   1   6   1   1   3   6   3
 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225
   1   3   1   1   4   5   3   3   5   3   3   3   3   5   1   3   2   4   3   6   1   5   6   5   3   4   2   5   5   1   5   3   2   3   3   5   6   4   6   5   3   1   3   3   5
 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 269 270 271
   1   1   1   1   1   5   1   1   1   1   5   2   5   3   5   1   5   5   1   1   2   1   5   5   1   5   1   2   1   5   5   1   3   5   1   5   3   1   4   1   1   6   5   4   1
 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 304 305 306 307 308 309 310 311 312 313 314 315 316 317
   3   1   5   2   2   3   2   3   2   1   3   2   5   3   2   3   4   3   5   4   3   3   5   1   3   3   1   3   1   5   4   6   3   2   4   3   6   3   2   5   2   1   3   5   3
 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363
   3   5   6   5   4   1   6   3   3   3   1   4   3   4   4   4   1   5   4   5   5   3   3   5   5   1   3   5   3   5   1   1   6   5   1   5   4   4   3   1   4   6   4   5   1
 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408
   5   3   3   5   5   5   3   2   5   1   1   3   5   5   3   6   1   5   5   5   1   5   3   5   6   5   5   3   5   5   5   5   1   1   1   5   5   5   1   6   1   4   3   5
 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453
   5   5   5   5   6   5   5   3   3   6   5   6   2   5   3   6   5   1   3   5   3   5   1   6   5   6   5   6   6   6   6   6   6   6   6   6   6   6   6   1   1   5   1   1
 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498
   3   1   5   3   3   3   3   2   6   3   6   4   3   3   3   6   3   1   6   6   3   5   5   5   4   6   5   5   2   6   6   5   5   6   4   5   1   3   6   5   5   3   5   1
 499 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544
   3   6   5   3   3   4   3   3   3   3   3   6   2   4   5   3   5   2   3   6   6   3   5   5   5   6   1   3   3   3   6   2   6   6   3   5   3   3   6   5   1   1   6   3   2
 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589
   5   1   6   6   2   1   6   6   2   1   3   6   3   3   3   5   3   1   6   5   4   1   6   1   3   6   5   1   4   4   5   1   5   1   5   5   3   2   4   3   5   1   3   6   6
 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634
   1   3   4   6   2   4   6   1   5   1   6   6   6   4   5   6   3   5   1   6   6   3   5   3   4   6   3   3   3   3   6   6   5   6   6   3   3   3   5   2   3   3   6
 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 671 672 673 674 675 676 677 678 679 681
   6   1   5   3   3   3   3   3   2   6   6   6   3   5   4   3   5   4   4   5   4   4   3   4   3   6   2   5   3   1   6   6   1   6   6   4   6   6   3   3   6   3   1   6
 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726
   4   5   1   1   3   5   5   1   3   4   2   3   6   5   3   3   3   3   5   6   5   3   2   6   6   3   3   5   5   3   5   2   3   5   5   2   5   5   5   2   5   5   3
 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771
   2   5   3   2   2   5   2   5   3   2   5   5   5   2   3   3   1   3   5   5   5   1   5   3   1   1   3   4   1   5   6   4   1   6   1   3   5   5   5   6   6   6   1   4   1
 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816
   3   4   6   2   3   6   2   1   5   5   3   5   3   5   5   5   4   5   5   6   5   5   2   5   5   1   1   3   4   5   1   3   5   5   3   3   3   3   1   6   6   6   6   2   6
 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861
   3   5   5   5   5   3   4   5   1   6   1   1   1   3   1   4   5   3   1   4   3   1   5   5   6   5   2   3   4   1   6   3   3   5   2   5   6   5   4   6   5   4   6
 862 863 864 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908
   5   3   5   5   3   3   5   6   5   6   5   3   2   5   6   5   1   3   3   3   6   6   4   3   3   6   1   3   5   5   4   3   1   3   5   4   2   5   2   2   5   5
 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953
   3   3   3   1   5   4   3   1   1   6   6   3   2   6   3   3   4   5   3   3   5   1   4   5   3   3   3   2   3   5   3   5   5   3   4   5   4   6   6   5   2   3   6   5   5
 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998
   6   2   3   5   3   5   3   3   3   3   5   5   6   3   6   6   5   5   3   6   3   4   1   1   4   3   3   3   1   4   4   3   4   4   5   6   2   6   5   5   3   5   5   5   5
```

c.

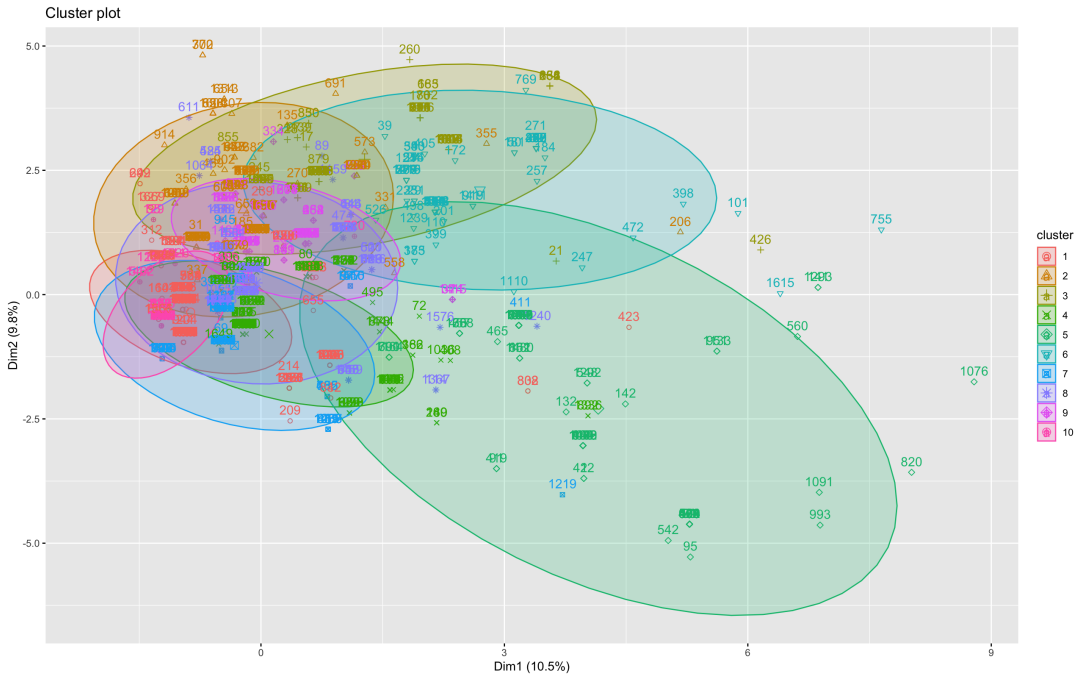
When k = 6, the graph as shown below is the categorization results in part b. Obviously, there are still a lot of clusters overlapping. We will increase the number of k slowly to find the optimal solution.



When k = 10, the WSS value is lowest , and the ration about WSS(Within Cluster Sum of Squares)/TSS(Total Cluster Sum of Squares) is closest to 0. So we decide k = 10 to be the optimal solution.



It is much better than the result when setting k = 6.



Check number of movies in every cluster and make sure that there is no one cluster will has too few movies.

	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8	cluster9	cluster10
number	529	151	84	304	71	68	130	165	78	84

When k = 6, the average of Cluster 6 is too low to be classified into any category. But when setting k = 10, We can observe obvious improvement. The number of movies which can not be categorized is decreased from 485 to 165, helping to make more movies be categorized precisely.