DATA ANALYSIS

Hierarchical cluster

Instructor: Prof. Christina Andresson

Presentors: Group 4

Our team

Vu Hoang Tuan Anh
VGU ID: 18812

Tran Kim Hoan
VGU ID: 18810

Ba Nguyen Quoc Anh
VGU ID: 17965

Nguyen Hoang Hai Nam
VGU ID: 17035

Content

01. Introduction

Objectives

method

Conduct a survey, collect data, analyse data inputs in R software by using hierarchical clustering and showcase insights.

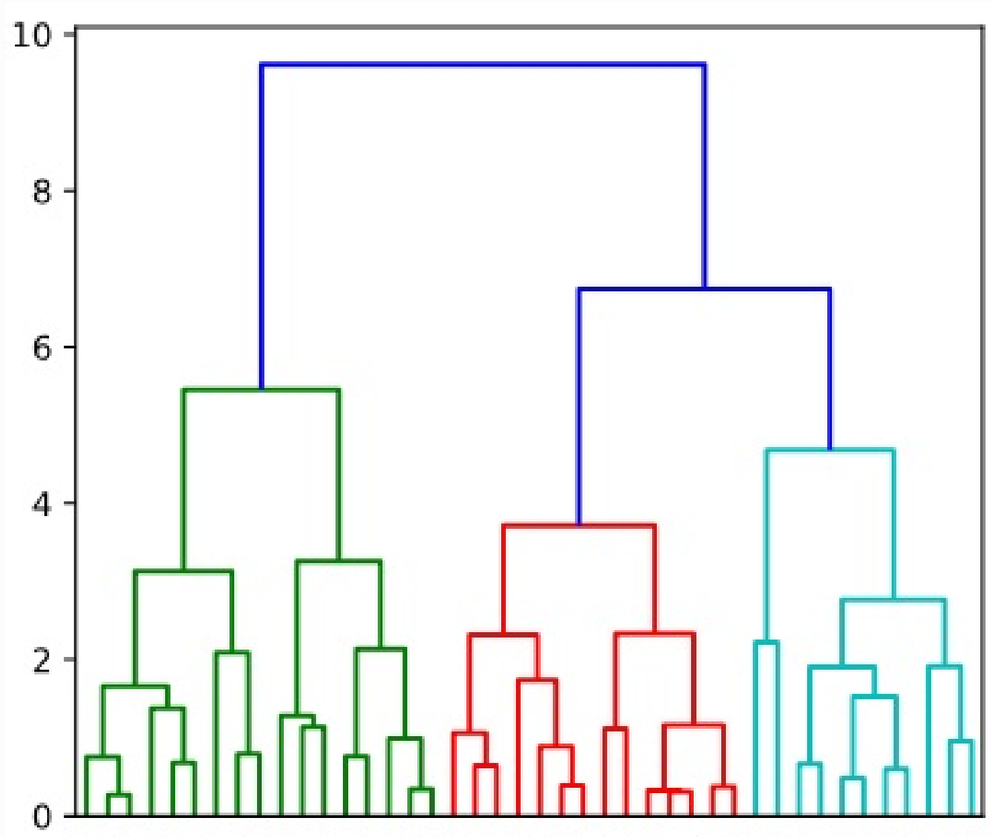
method

Inputting data into the analysis model, visualizing the results with a dendrogram

Objective 03

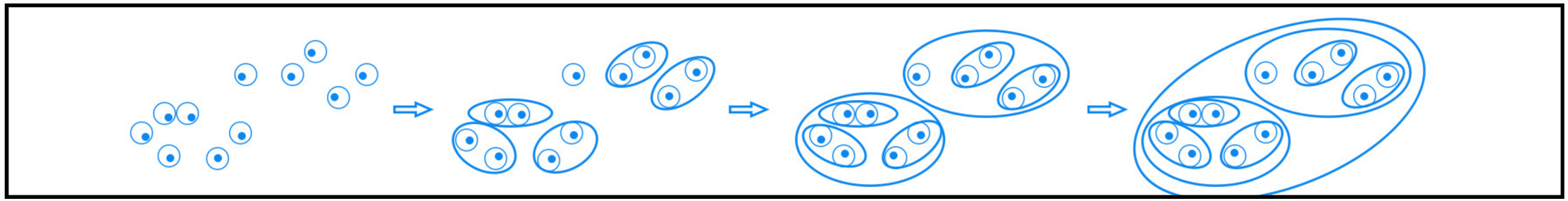
Grouping users with similar viewing patterns to recommend relevant genres based on survey questions.

What is “Hierarchical cluster analysis”?

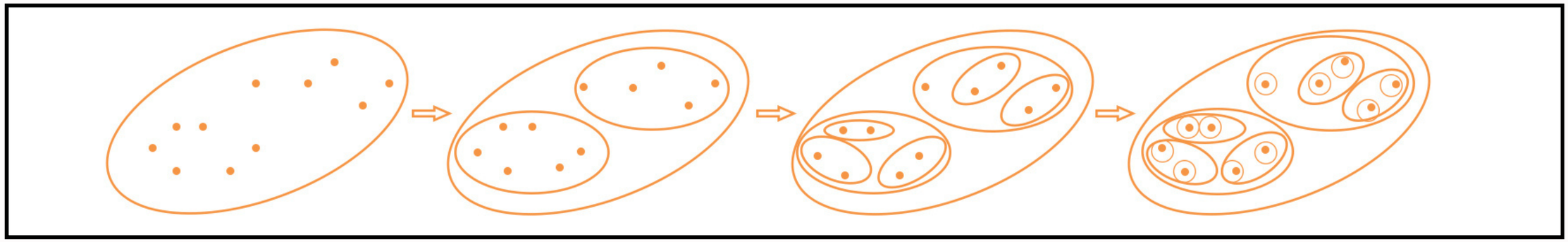


Hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters

Agglomerative



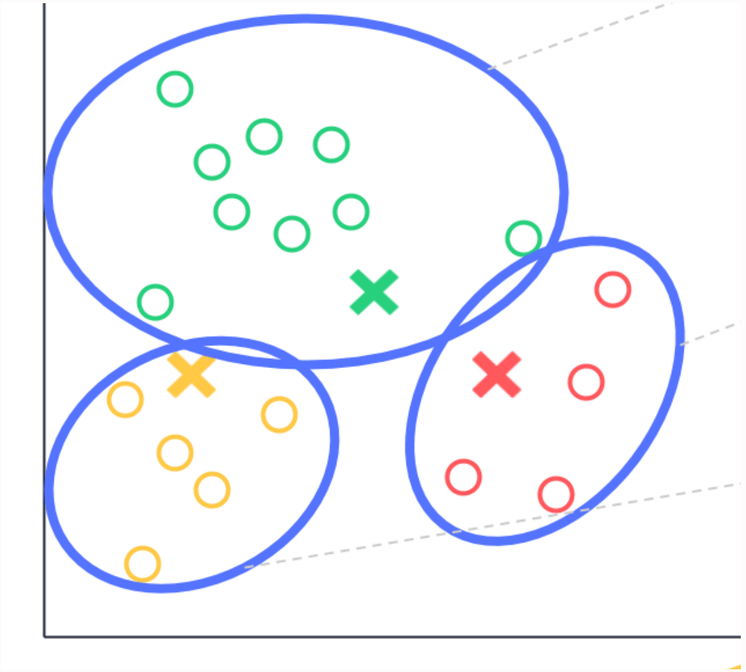
Diversive



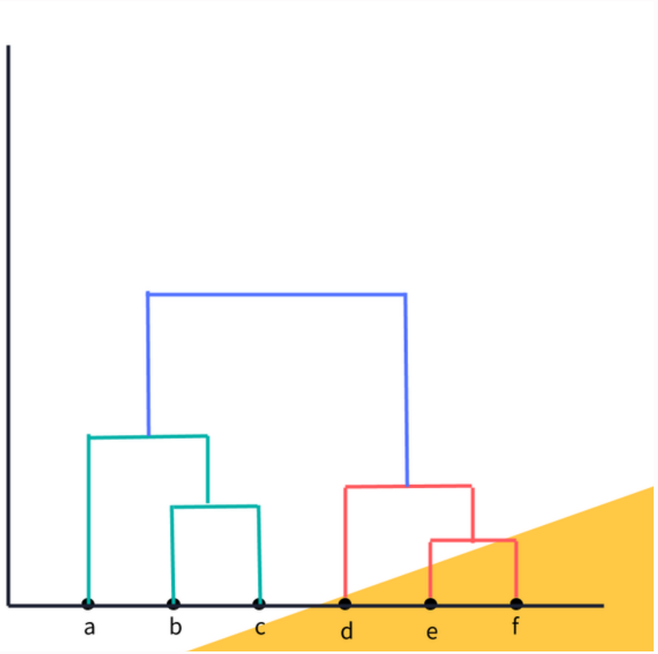
How does “hierarchical cluster analysis” work?

Find distance

Grouping



Termination



02. Data

Data is derived from an online survey via Google form

Step 3
Data Preparation

Step 2
Data Collection

Step 1
Attributes Identification and Selection

Attributes Identification & Selection: “Movie Genres Preference”

7 key attributes

1. Age
2. Gender
3. Working/learning area
4. Preferred film genre
5. Factor influencing genre choice
6. Frequency of movie watching
7. Source of film viewing

Data Collection

Create a survey with 7 questions

1. How old are you?
2. What is your gender?
3. Which area do you work/learn in?
4. Which movie genre is your favorite?
5. How often do you watch movies per week?
6. What is the factor that influences your decision to watch a movie?
7. Which platform do you use to watch movies?

Movies Preferences Survey (Khảo sát sở thích xem phim)

This form is used to survey your movie preferences. Please answer the following questions to help us understand your movie-watching habits.
Biểu mẫu này được dùng để khảo sát sở thích xem phim. Câu trả lời của bạn sẽ giúp chúng tôi hiểu rõ về sở thích xem phim của bạn.

* Indicates required question

01. How old are you? (Eg. 18) (Bạn bao nhiêu tuổi? (Ví dụ 18)) *

Your answer

02. What is your gender? (Giới tính của bạn là gì?) *

- Male (Nam)
- Female (Nữ)
- Other (Khác)

11/5/23, 10:00 PM Movies Preferences Survey (Khảo sát sở thích xem phim)

03. Which area do you work/learn in? (Bạn làm việc/học tập trong lĩnh vực nào?) *

- Accounting/Finance/Sales/Marketing (Kế toán/Tài chính/Sales/Marketing)
- Engineering (Kĩ thuật)
- Education (Giáo dục)
- Information Technology and related fields (Công nghệ thông tin và các lĩnh vực liên quan)
- Arts/Entertainment (Nghệ thuật/Giải trí)
- Government/Non-profit (Chính phủ/Phi chính phủ)
- Society: Journalism/Law/Languages... (Xã hội: Báo chí/Luật/Ngôn ngữ...)
- Healthcare (Chăm sóc sức khỏe)
- Agriculture (Nông nghiệp)
- Other:

04. Which movie genre is your favorite? ***Choose only 1*** *
(Thể loại phim yêu thích nhất của bạn là gì?) ***Chỉ chọn 1***

- Action (Hành động)
- Animation/Anime/Cartoon (Hoạt hình)
- Comedy (Hài)
- Documentary/History (Tài liệu/Lịch sử)
- Drama (Chính kịch)
- Horror (Kinh dị)
- Science-fiction (Khoa học viễn tưởng)
- Fantasy (Kỳ ảo)
- Romance (Lãng mạn)
- Other:

Survey Image 1

05. What is the factor that influences your decision to watch a movie? (Yếu tố nào ảnh hưởng đến quyết định của bạn khi xem 1 bộ phim?) *

- ◯ The title and description of the movie (Tựa đề và phần mô tả phim)
- ◯ The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu bộ phim)
- ◯ The cast and crew of the movie (Dàn diễn viên và đoàn làm phim)
- ◯ The duration of the movie (Thời lượng của bộ phim)
- ◯ The quality of the movie (Chất lượng của bộ phim)
- ◯ Ratings of the movie (Đánh giá của bộ phim)
- ◯ Random choice (Chọn ngẫu nhiên)
- ◯ Other:

06. How often do you watch movies **per week**? (Tần suất bạn xem phim **mỗi tuần**?) *

- ◯ Less than 2 hours (Ít hơn 2 giờ)
- ◯ 2 - 5 hours (2 - 5 giờ)
- ◯ 6 - 10 hours (6 - 10 giờ)
- ◯ 11 - 15 hours (11 - 15 giờ)
- ◯ 16 - 20 hours (16 - 20 giờ)
- ◯ More than 20 hours (Nhiều hơn 20 giờ)

11/5/23, 10:00 PM Movies Preferences Survey (Khảo sát sở thích xem phim)

07. Which platform do you use to watch movies?
(Bạn sử dụng nền tảng nào để xem phim?)

- ◯ Television (Ti-vi)
- ◯ Mobile app: Netflix, WeTV, VieON, IQIYI,... (Ứng dụng di động: Netfix, WeTV, VieON, IQIYI,...)
- ◯ Movie theater (Rạp chiếu phim)
- ◯ Web browser (Trình duyệt mạng Internet)
- ◯ DVDs/Blu-rays (Đĩa cứng)
- ◯ Other:

Submit Page 1 of 1 Clear form

Survey Image 2

Individuals **aged 16 and above**

A total of **118 responses** from survey

	A	B	C	D	E	F	G	
1	Timestamp	01. How old are you? ((Bạn bao nhiêu tuổi? (	02. What is your gender? (Giới tính của bạn là gì?)	03. Which area do you work/learn in? (Bạn làm việc/học tập trong lĩnh vực	04. Which movie genre is your favorite? (Thể loại phim yêu thích nhất của bạn là	06. How often do you watch movies per week? (Tần suất bạn xem phim mỗi tuần?)	05. What is the factor that influences your decision to watch a movie? (Yếu tố nào ảnh hưởng đến quyết định của bạn khi xem 1 bộ phim?)	07. Which platform do you u (Bạn sử dụng nền tảng nào
2	10/31/2023 11:15:05	21	Female (Nữ)	Accounting/Finance/Sales/Marketing	Horror (Kinh dị)	Less than 2 hours (Ít hơn 2 giờ)	Random choice (Chọn ngẫu nhiên)	Web browser (Trình duyệt n
3	10/31/2023 11:21:19	21	Male (Nam)	Engineering (Kĩ thuật)	Animation/Anime/Cartoon (Hoạt hình)	2 - 5 hours (2 - 5 giờ)	The title and description of the movie (Tựa đề và phần mô tả phim)	Web browser (Trình duyệt n
4	10/31/2023 11:27:54	17	Female (Nữ)	Society: Journalism/Law/Languages..	Animation/Anime/Cartoon (Hoạt hình)	6 - 10 hours (6 - 10 giờ)	Ratings of the movie (Đánh giá của bộ phim)	Television (Ti-vi)
5	10/31/2023 11:29:12	54	Male (Nam)	Agriculture (Nông nghiệp)	Action (Hành động)	6 - 10 hours (6 - 10 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Mobile app: Netflix, WeTV,
6	10/31/2023 11:30:06	19	Male (Nam)	Arts/Entertainment (Nghệ thuật/Giải t	Animation/Anime/Cartoon (Hoạt hình)	More than 20 hours (Nhiều hơn 20 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Web browser (Trình duyệt n
7	10/31/2023 11:30:56	26	Female (Nữ)	Society: Journalism/Law/Languages..	Action (Hành động)	11 - 15 hours (11 - 15 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Web browser (Trình duyệt n
8	10/31/2023 11:33:00	54	Female (Nữ)	Education (Giáo dục)	Romance (Lãng mạn)	11 - 15 hours (11 - 15 giờ)	The title and description of the movie (Tựa đề và phần mô tả phim)	Mobile app: Netflix, WeTV,
9	10/31/2023 11:35:10	21	Male (Nam)	Engineering (Kĩ thuật)	Science-fiction (Khoa học viễn tưởng)	2 - 5 hours (2 - 5 giờ)	The quality of the movie (Chất lượng của bộ phim)	Web browser (Trình duyệt n
10	10/31/2023 11:47:23	21	Male (Nam)	Information Technology and related fi	Crime	Less than 2 hours (Ít hơn 2 giờ)	Ratings of the movie (Đánh giá của bộ phim)	DVDs/Blu-rays (Đĩa cứng)
11	10/31/2023 11:55:31	21	Male (Nam)	Engineering (Kĩ thuật)	Science-fiction (Khoa học viễn tưởng)	Less than 2 hours (Ít hơn 2 giờ)	The quality of the movie (Chất lượng của bộ phim)	Web browser (Trình duyệt n
12	10/31/2023 11:56:29	21	Female (Nữ)	Accounting/Finance/Sales/Marketing	Comedy (Hài)	2 - 5 hours (2 - 5 giờ)	The title and description of the movie (Tựa đề và phần mô tả phim)	Movie theater (Rạp chiếu ph
13	10/31/2023 12:03:28	21	Female (Nữ)	Arts/Entertainment (Nghệ thuật/Giải t	Comedy (Hài)	More than 20 hours (Nhiều hơn 20 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Mobile app: Netflix, WeTV,
14	10/31/2023 12:13:20	21	Female (Nữ)	Education (Giáo dục)	Horror (Kinh dị)	6 - 10 hours (6 - 10 giờ)	The quality of the movie (Chất lượng của bộ phim)	Mobile app: Netflix, WeTV,
15	10/31/2023 12:13:50	21	Male (Nam)	Engineering (Kĩ thuật)	Science-fiction (Khoa học viễn tưởng)	2 - 5 hours (2 - 5 giờ)	Random choice (Chọn ngẫu nhiên)	Web browser (Trình duyệt n
16	10/31/2023 12:14:25	21	Female (Nữ)	Arts/Entertainment (Nghệ thuật/Giải t	Comedy (Hài)	Less than 2 hours (Ít hơn 2 giờ)	The cast and crew of the movie (Dàn diễn viên và đoàn làm phim)	Movie theater (Rạp chiếu ph
17	10/31/2023 12:14:55	20	Female (Nữ)	Accounting/Finance/Sales/Marketing	Romance (Lãng mạn)	6 - 10 hours (6 - 10 giờ)	The cast and crew of the movie (Dàn diễn viên và đoàn làm phim)	Mobile app: Netflix, WeTV,
18	10/31/2023 12:16:42	21	Female (Nữ)	Society: Journalism/Law/Languages..	Comedy (Hài)	Less than 2 hours (Ít hơn 2 giờ)	The cast and crew of the movie (Dàn diễn viên và đoàn làm phim)	Web browser (Trình duyệt n
19	10/31/2023 12:31:31	22	Male (Nam)	Engineering (Kĩ thuật)	Action (Hành động)	2 - 5 hours (2 - 5 giờ)	The quality of the movie (Chất lượng của bộ phim)	Mobile app: Netflix, WeTV,
20	10/31/2023 12:36:33	21	Male (Nam)	Information Technology and related fi	Action (Hành động)	Less than 2 hours (Ít hơn 2 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Web browser (Trình duyệt n
21	10/31/2023 12:37:09	28	Other (Khác)	Arts/Entertainment (Nghệ thuật/Giải t	Fantasy (Kỳ ảo)	11 - 15 hours (11 - 15 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Mobile app: Netflix, WeTV,
22	10/31/2023 12:38:59	21	Female (Nữ)	Engineering (Kĩ thuật)	Horror (Kinh dị)	11 - 15 hours (11 - 15 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Mobile app: Netflix, WeTV,
23	10/31/2023 12:43:08	17	Male (Nam)	Accounting/Finance/Sales/Marketing	Romance (Lãng mạn)	Less than 2 hours (Ít hơn 2 giờ)	The cast and crew of the movie (Dàn diễn viên và đoàn làm phim)	Mobile app: Netflix, WeTV,
24	10/31/2023 12:45:40	21	Female (Nữ)	Graphic Design	Fantasy (Kỳ ảo)	2 - 5 hours (2 - 5 giờ)	Ratings of the movie (Đánh giá của bộ phim)	Web browser (Trình duyệt n
25	10/31/2023 12:47:06	20	Male (Nam)	Tourism	Animation/Anime/Cartoon (Hoạt hình)	6 - 10 hours (6 - 10 giờ)	The quality of the movie (Chất lượng của bộ phim)	Web browser (Trình duyệt n
26	10/31/2023 12:49:06	33	Female (Nữ)	Education (Giáo dục)	Fantasy (Kỳ ảo)	2 - 5 hours (2 - 5 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Mobile app: Netflix, WeTV,
27	10/31/2023 12:55:00	22	Male (Nam)	Information Technology and related fi	Science-fiction (Khoa học viễn tưởng)	Less than 2 hours (Ít hơn 2 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Mobile app: Netflix, WeTV,
28	10/31/2023 12:57:33	22	Male (Nam)	Information Technology and related fi	Comedy (Hài)	2 - 5 hours (2 - 5 giờ)	The quality of the movie (Chất lượng của bộ phim)	Web browser (Trình duyệt n
29	10/31/2023 12:59:13	I'm 16.	Female (Nữ)	Society: Journalism/Law/Languages..	Comedy (Hài)	Less than 2 hours (Ít hơn 2 giờ)	The image and trailer of the movie (Hình ảnh và đoạn phim giới thiệu	Web browser (Trình duyệt n
30	10/31/2023 13:02:08	21	Male (Nam)	Arts/Entertainment (Nghệ thuật/Giải t	Documentary/History (Tài liệu/Lịch sử)	2 - 5 hours (2 - 5 giờ)	The title and description of the movie (Tựa đề và phần mô tả phim)	Mobile app: Netflix, WeTV,
31	10/31/2023 13:03:41	16	Female (Nữ)	High school student	Comedy (Hài)	Less than 2 hours (Ít hơn 2 giờ)	The quality of the movie (Chất lượng của bộ phim)	Mobile app: Netflix, WeTV,
32	10/31/2023 13:09:33	16	Female (Nữ)	Accounting/Finance/Sales/Marketing	Comedy (Hài)	Less than 2 hours (Ít hơn 2 giờ)	The quality of the movie (Chất lượng của bộ phim)	Mobile app: Netflix, WeTV,

Data Preparation: A crucial step

Data

Contains errors, outliers, inconsistent, missing values, etc.
In wrong formats

Efficiency

- Clean data
- Change to standardized formats

Better results

Change from question format to words/phrases format:

01. How old are you? (Eg. 18) (Bạn bao nhiêu tuổi? (Ví dụ 18))	Age
02. What is your gender? (Giới tính của bạn là gì?)	Gender
03. Which area do you work/learn in? (Bạn làm việc/học tập trong lĩnh vực nào?)	Field
04. Which movie genre is your favorite? Choose only 1 (Thể loại phim yêu thích nhất của bạn là gì?) Chỉ chọn 1	Genre
05. How often do you watch movies per week? (Tần suất bạn xem phim mỗi tuần?)	Frequency
06. What is the factor that influences your decision to watch a movie? (Yếu tố nào ảnh hưởng đến quyết định của bạn khi xem 1 bộ phim?)	Factor
07. Which platform do you use to watch movies? (Bạn sử dụng nền tảng nào để xem phim?)	Platform

Substitute long words/phrases with shorter words/phrases:

Accounting/Finance/Sales/Marketing (Kế toán/Tài chính/Sales/Marketing)	Business
Engineering (Kĩ thuật)	Engineering
Society: Journalism/Law/Languages... (Xã hội: Báo chí/Luật/Ngôn ngữ...)	Society
Agriculture (Nông nghiệp)	Agriculture
Arts/Entertainment (Nghệ thuật/Giải trí)	Arts/Entertainment
Society: Journalism/Law/Languages... (Xã hội: Báo chí/Luật/Ngôn ngữ...)	Society
Education (Giáo dục)	Education
Engineering (Kĩ thuật)	Engineering
Information Technology and related fields (Công nghệ thông tin và các lĩnh vực liên quan)	Information Technology
Engineering (Kĩ thuật)	Engineering
Accounting/Finance/Sales/Marketing (Kế toán/Tài chính/Sales/Marketing)	Business

Change from wrong format to standardized format:

Age	Gender
22	Male
I'm 16.	Female
21	Male
16	Female

Age	Gender
22	Male
16	Female
21	Male
16	Female

Eliminate responses with unexpected targetted objects:

28	Female	Agriculture
yeah im 12	Male	im a student
21	Male	Education

28	Female	Agriculture
21	Male	Education

Final data with **114 responses:**

Age	Gender	Field	Genre	Frequency	Factor	Platform
16	Female	Society	Comedy	Less than 2 hours	The image and trailer of the movie	Web browser
16	Female	Business	Comedy	Less than 2 hours	The quality of the movie	Mobile app
17	Female	Society	Animation	6 - 10 hours	Ratings of the movie	Television
17	Male	Business	Romance	Less than 2 hours	The cast and crew of the movie	Mobile app
18	Female	Society	Action	2 - 5 hours	The quality of the movie	Web browser
18	Male	Arts/Entertainment	Horror	Less than 2 hours	The cast and crew of the movie	Movie theater
18	Female	Education	Fantasy	Less than 2 hours	The quality of the movie	Mobile app
19	Male	Arts/Entertainment	Animation	More than 20 hours	The image and trailer of the movie	Web browser
19	Male	Information Technology	Documentary/History	Less than 2 hours	The title and description of the movie	Mobile app
19	Female	Engineering	Romance	2 - 5 hours	The title and description of the movie	Mobile app
19	Male	Information Technology	Science-fiction	Less than 2 hours	The cast and crew of the movie	Movie theater
19	Male	Information Technology	Action	More than 20 hours	The image and trailer of the movie	Mobile app
19	Male	Engineering	Animation	More than 20 hours	The quality of the movie	Web browser
19	Female	Information Technology	Action	2 - 5 hours	The title and description of the movie	Mobile app
20	Female	Business	Romance	6 - 10 hours	The cast and crew of the movie	Mobile app

03. Hierarchical Methodology and Result

Hierarchical clustering methodology

In this course, we use R programming language
to analysis and R Studio as an IDE

version 4.2.1

2023.09.1 Build 494

Hierarchical clustering methodology

- To start applying the Hierarchical clustering model, we import the data file and store it as a data frame in R

```
# Read the data frame
df <- read.csv("data/clean_data_v2.csv")

# Omit the NA values of the data frame
df <- na.omit(df)
```

	Age	Gender	Field	Genre	Frequency	Factor	Platform
1	16	Female	Society	Comedy	Less than 2 hours	The image and trailer of the movie	Web browser
2	16	Female	Business	Comedy	Less than 2 hours	The quality of the movie	Mobile app
3	17	Female	Society	Animation	6 - 10 hours	Ratings of the movie	Television
4	17	Male	Business	Romance	Less than 2 hours	The cast and crew of the movie	Mobile app
5	18	Female	Society	Action	2 - 5 hours	The quality of the movie	Web browser
6	18	Male	Arts/Entertainment	Horror	Less than 2 hours	The cast and crew of the movie	Movie theater
7	18	Female	Education	Fantasy	Less than 2 hours	The quality of the movie	Mobile app
8	19	Male	Arts/Entertainment	Animation	More than 20 hours	The image and trailer of the movie	Web browser
9	19	Male	Information Technology	Documentary/History	Less than 2 hours	The title and description of the movie	Mobile app

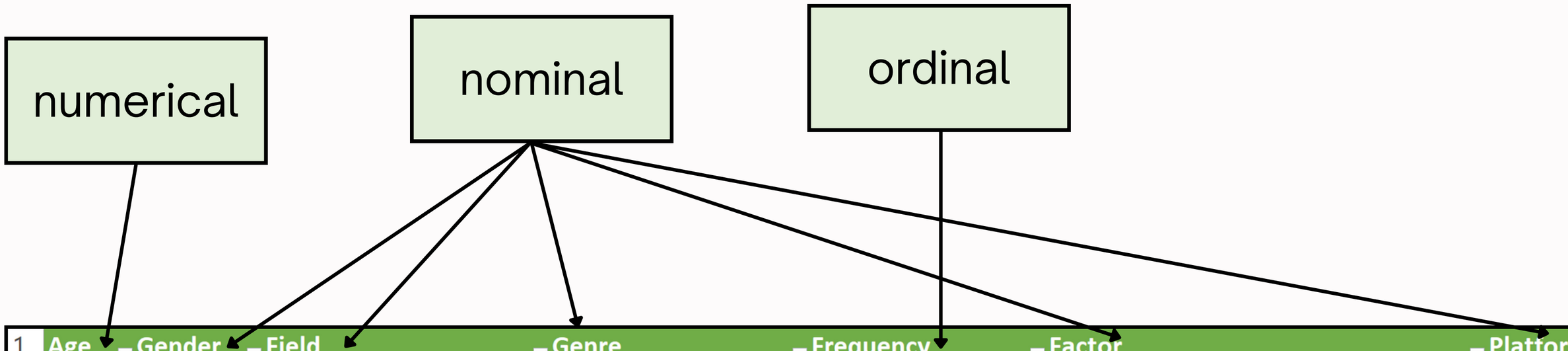
Hierarchical clustering methodology

- Then we standardize (scale) the Age (numerical variable) and Frequency (ordinal variable)

```
# Standardize the Age variable
df$Age_std <- scale(df$Age)

# Standardize the Frequency variable
df$Frequency <- factor(
      df$Frequency,
      order = TRUE,
      levels = c("Less than 2 hours", "2 - 5 hours", "6 - 10 hours",
"11 - 15 hours", "16 - 20 hours", "More than 20 hours"))
df$Frequency_numeric <- as.numeric(factor(df$Frequency))
df$Frequency_std <- scale(df$Frequency_numeric)
```

Hierarchical clustering methodology

- The data frame consists of categorical, numerical and ordinal variables



	Age	Gender	Field	Genre	Frequency	Factor	Platform
2	16	Female	Society	Comedy	Less than 2 hours	The image and trailer of the movie	Web browser
3	16	Female	Business	Comedy	Less than 2 hours	The quality of the movie	Mobile app
4	17	Female	Society	Animation	6 - 10 hours	Ratings of the movie	Television
5	17	Male	Business	Romance	Less than 2 hours	The cast and crew of the movie	Mobile app
6	18	Female	Society	Action	2 - 5 hours	The quality of the movie	Web browser
7	18	Male	Arts/Entertainment	Horror	Less than 2 hours	The cast and crew of the movie	Movie theater
8	18	Female	Education	Fantasy	Less than 2 hours	The quality of the movie	Mobile app

Hierarchical clustering methodology

- We can not calculate the distance between 2 entries of the data frame by using mathematical algorithms (Euclidean, Manhattan, Canberra, etc) directly since all variables are not numerical.

1	Age	Gender	Field	Genre	Frequency	Factor	Platform
2	16	Female	Society	Comedy	Less than 2 hours	The image and trailer of the movie	Web browser
3	16	Female	Business	Comedy	Less than 2 hours	The quality of the movie	Mobile app
4	17	Female	Society	Animation	6 - 10 hours	Ratings of the movie	Television
5	17	Male	Business	Romance	Less than 2 hours	The cast and crew of the movie	Mobile app
6	18	Female	Society	Action	2 - 5 hours	The quality of the movie	Web browser
7	18	Male	Arts/Entertainment	Horror	Less than 2 hours	The cast and crew of the movie	Movie theater
8	18	Female	Education	Fantasy	Less than 2 hours	The quality of the movie	Mobile app

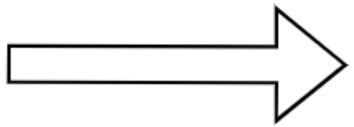
Hierarchical clustering methodology

- **Solution:** Using Dummy variables

```
# Turn the categorical variables into dummy variables
df_dummy <- model.matrix(~ Age_std + Genre + Frequency_std + Field + Factor + Gender +
Platform, data = df)
```

Genre
Action
Comedy
Fantasy

Dummy variables



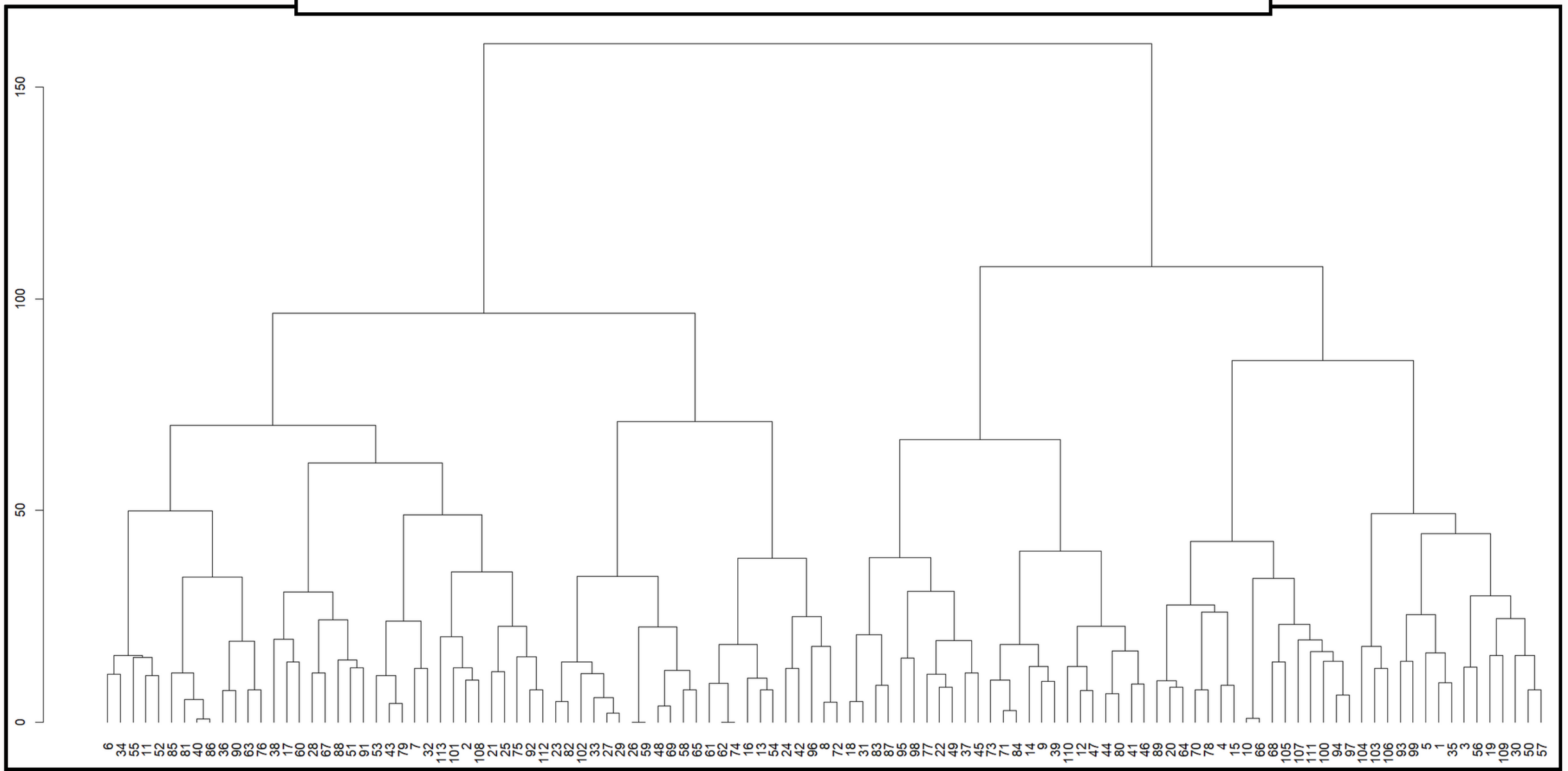
	Genre.Action	Genre.Comedy
Action	1	0
Comedy	0	1
Fantasy	0	0

Hierarchical clustering methodology

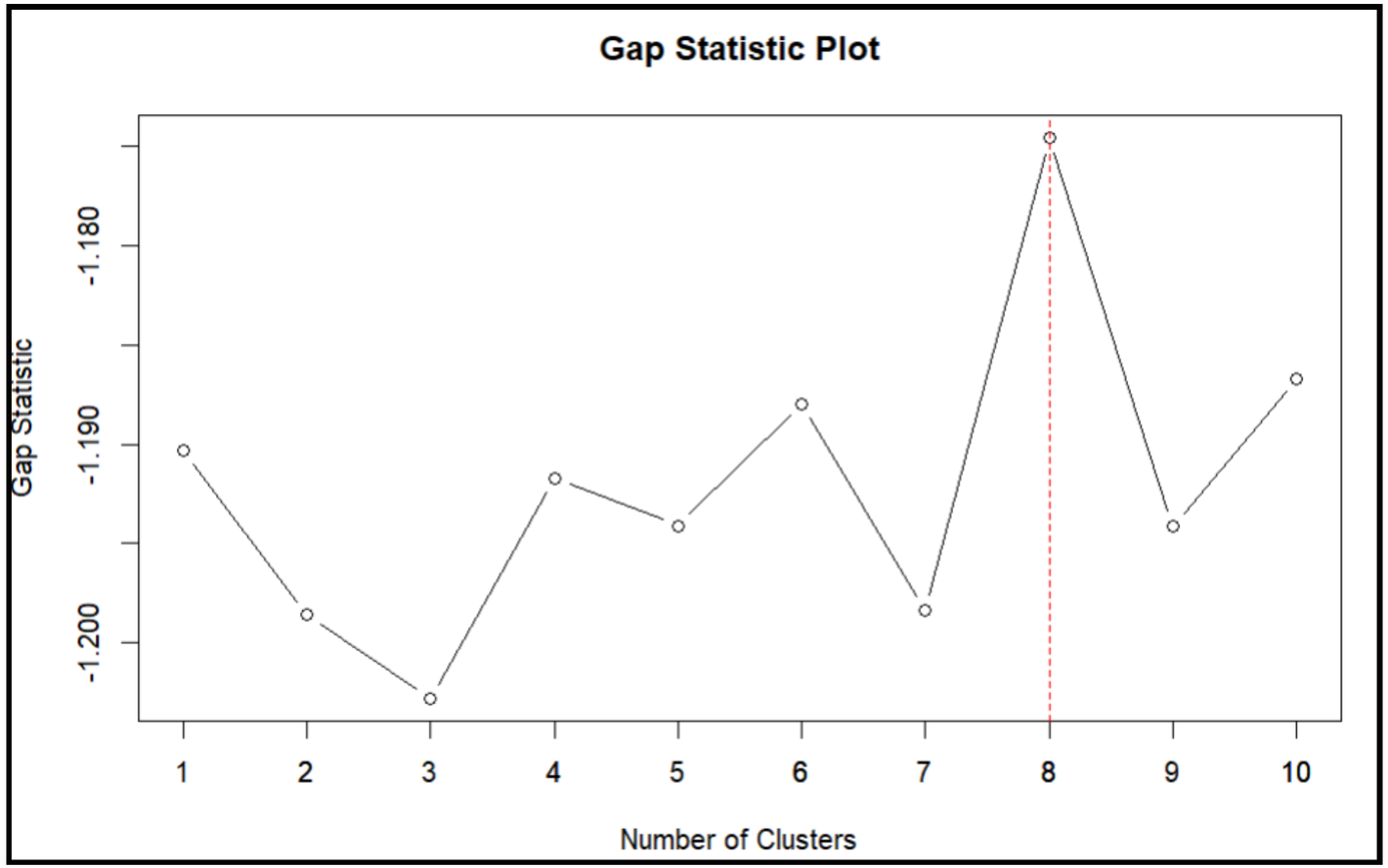
- **Now we calculate the points distance vector and then the cluster distances**
  - **Points distance measures:** Euclidean, Manhattan, Canberra, Maximum, Binary, Minkowski, etc
  - **Cluster distance measures:** Ward.D, Single, Complete, Average, Centroid, etc

```
# Calculate the points distance
point_dist <- dist(df_dummy, method = "canberra")       # Using Canberra distance
# Hierarchical cluster analysis on the data frame
hc <- hclust(point_dist, method = "ward.D")             # Using Ward's method
```

```r
# Plot the dendrogram
dend <- as.dendrogram(hc)
plot(dend)
```



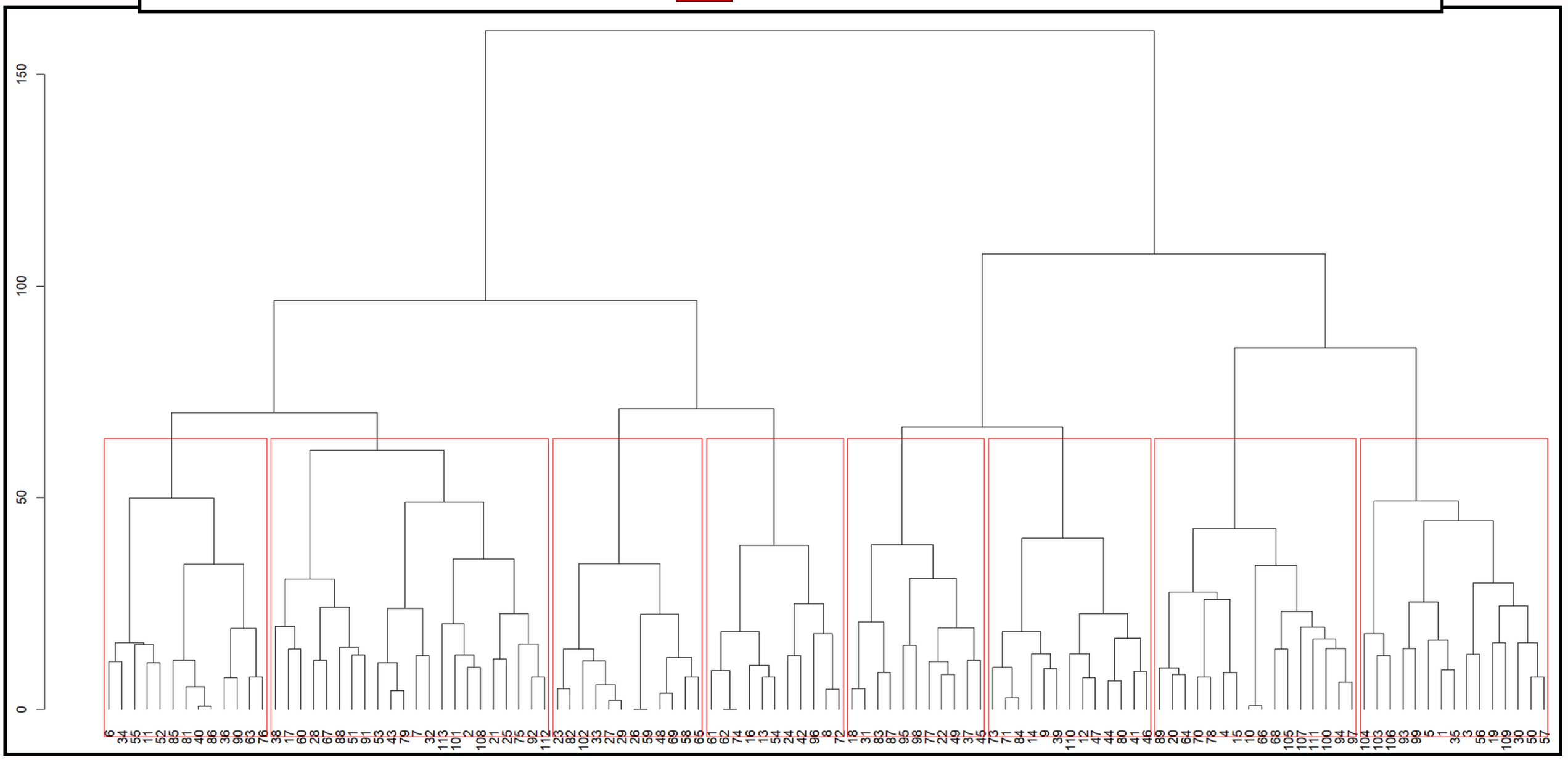
Finding optimal numbers of clusters



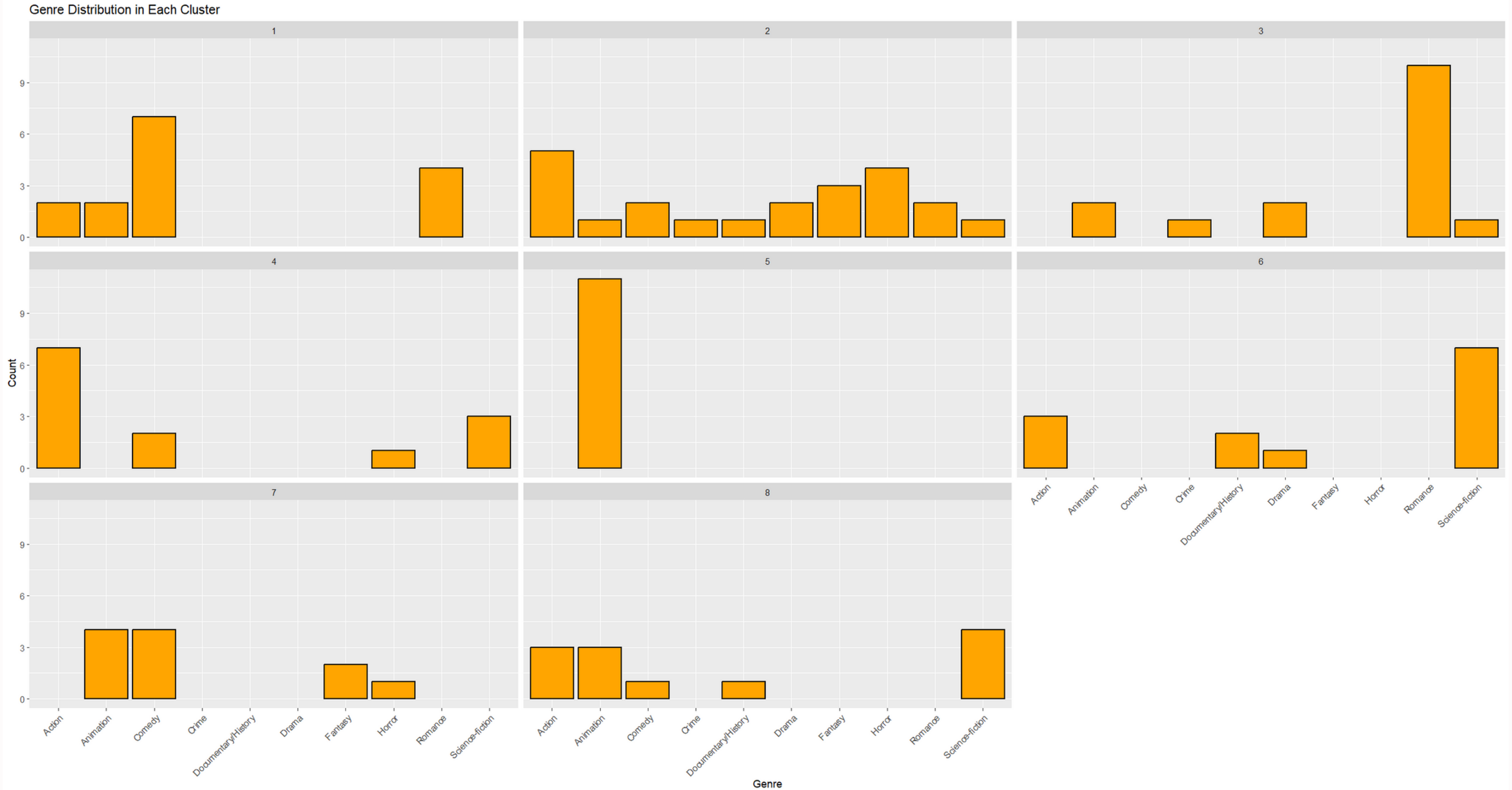
Gap Statistic Method

- This approach can be applied to any clustering method.
- The estimate of the optimal clusters will be the value that maximizes the gap statistic. This means that the clustering structure is far away from the random uniform distribution of points.

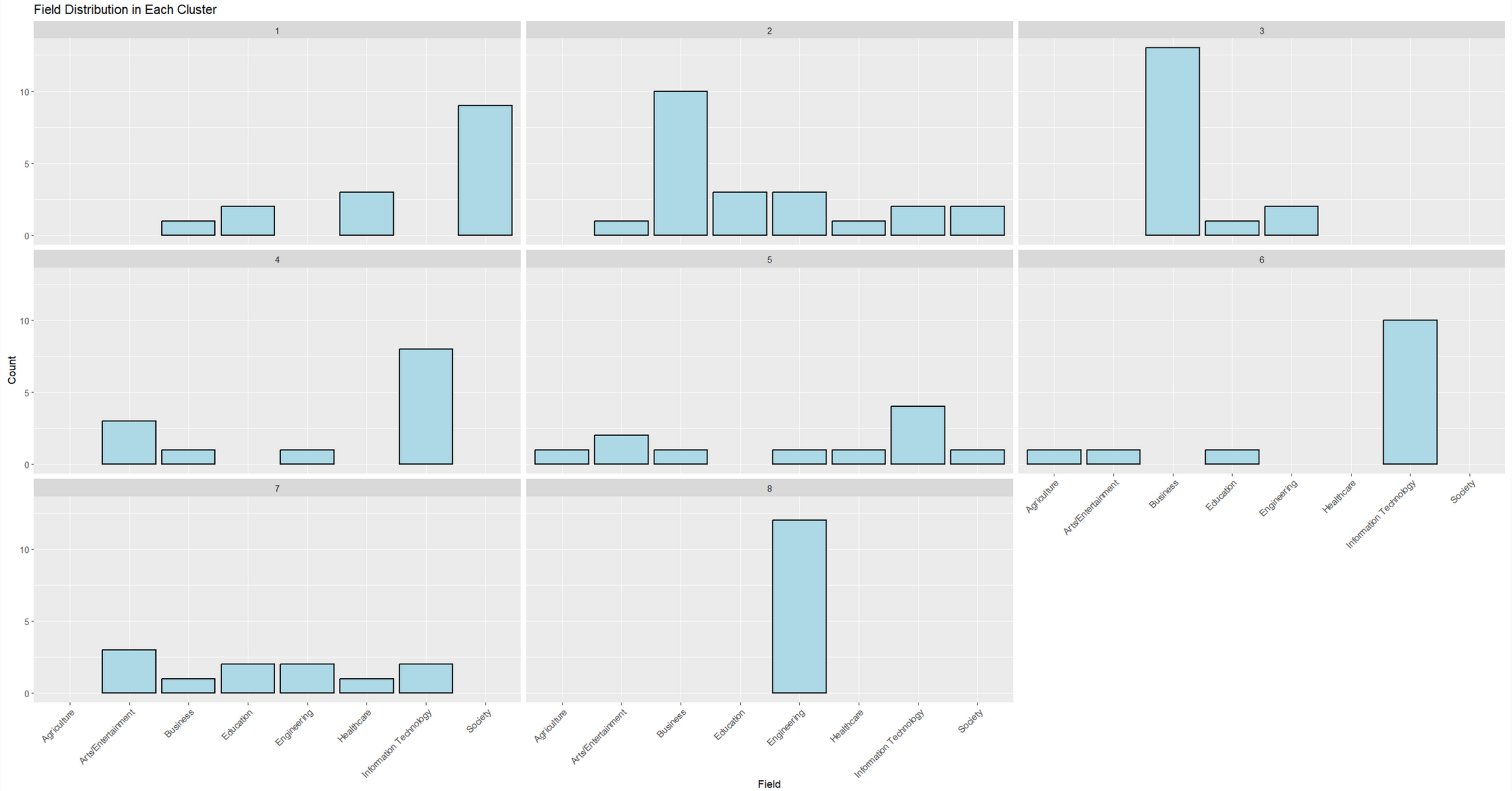
```r
# Draw the rectangle around each cluster in k clusters
k <- 8
rect.hclust(hc, k, border = "red")
```

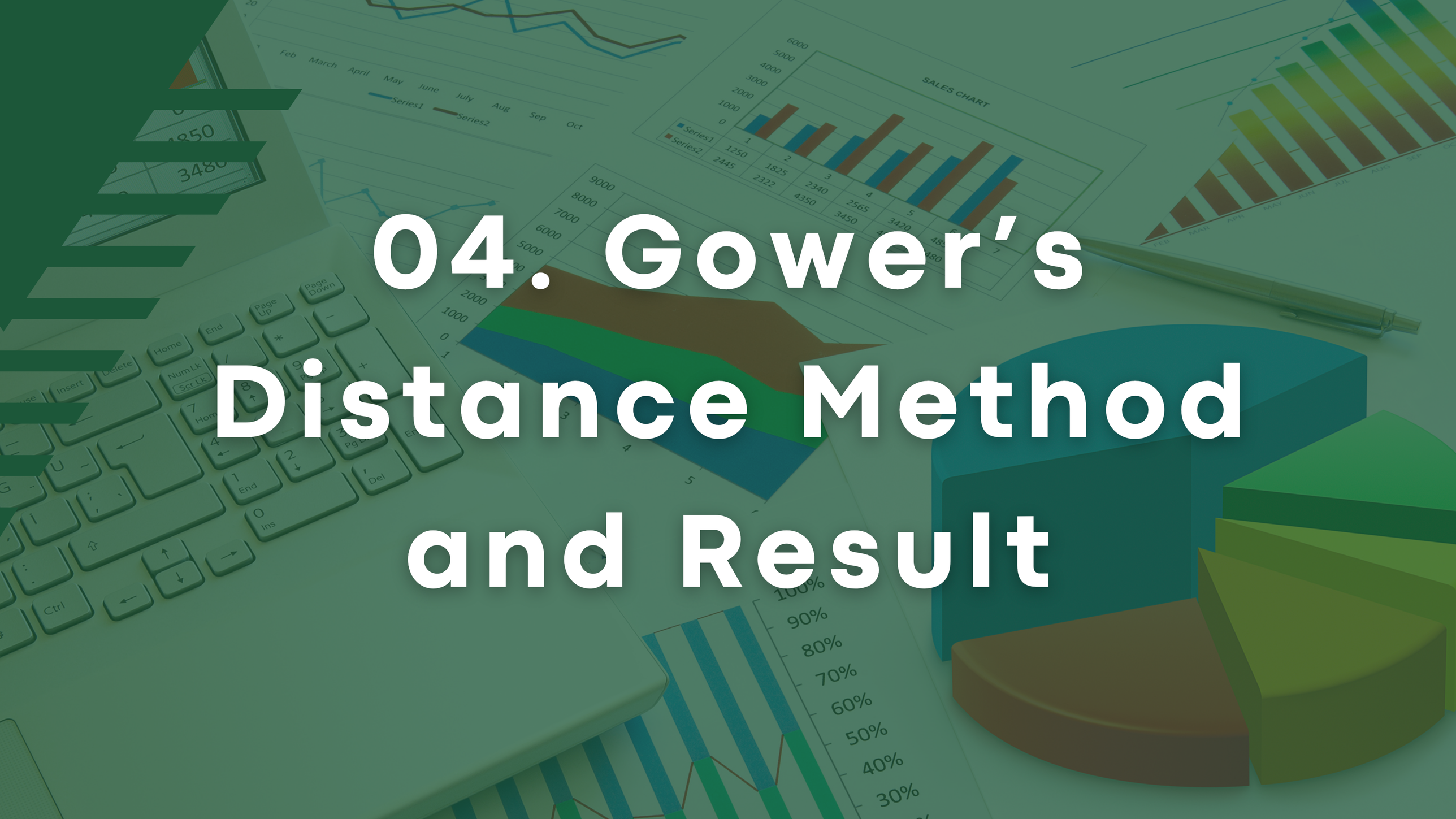


Genres Distribution in each cluster



Fields Distribution in each cluster





04. Gower's Distance Method and Result

Gower’s Distance

Definition

Distance = (Dist[Age] + Dist[Factor] + ...) / “Number of attributes”

Types of Data

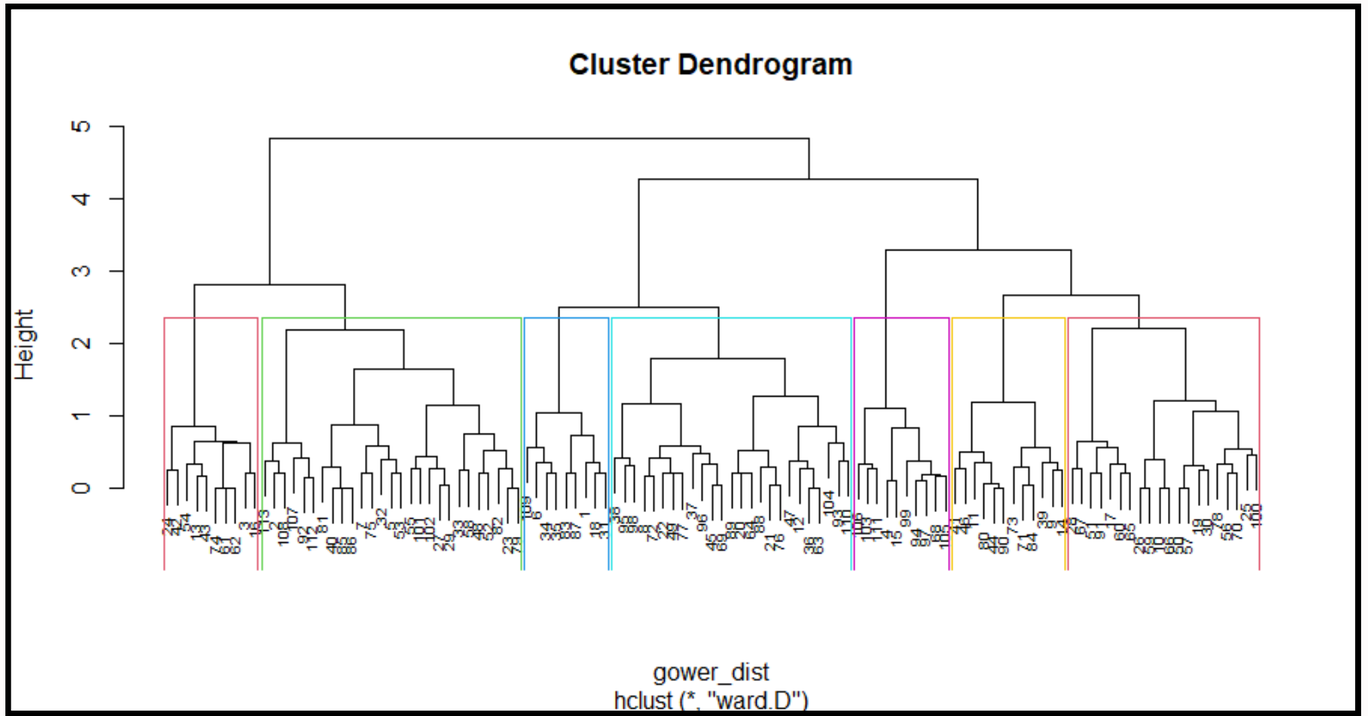
+ Numerical: Absolute value of the distance divided by the range.
+ Ordinal: Same as numerical with converted ordinal values.
+ Categorical:
  - Identical --> 0
  - Different --> 1

Data Preparation

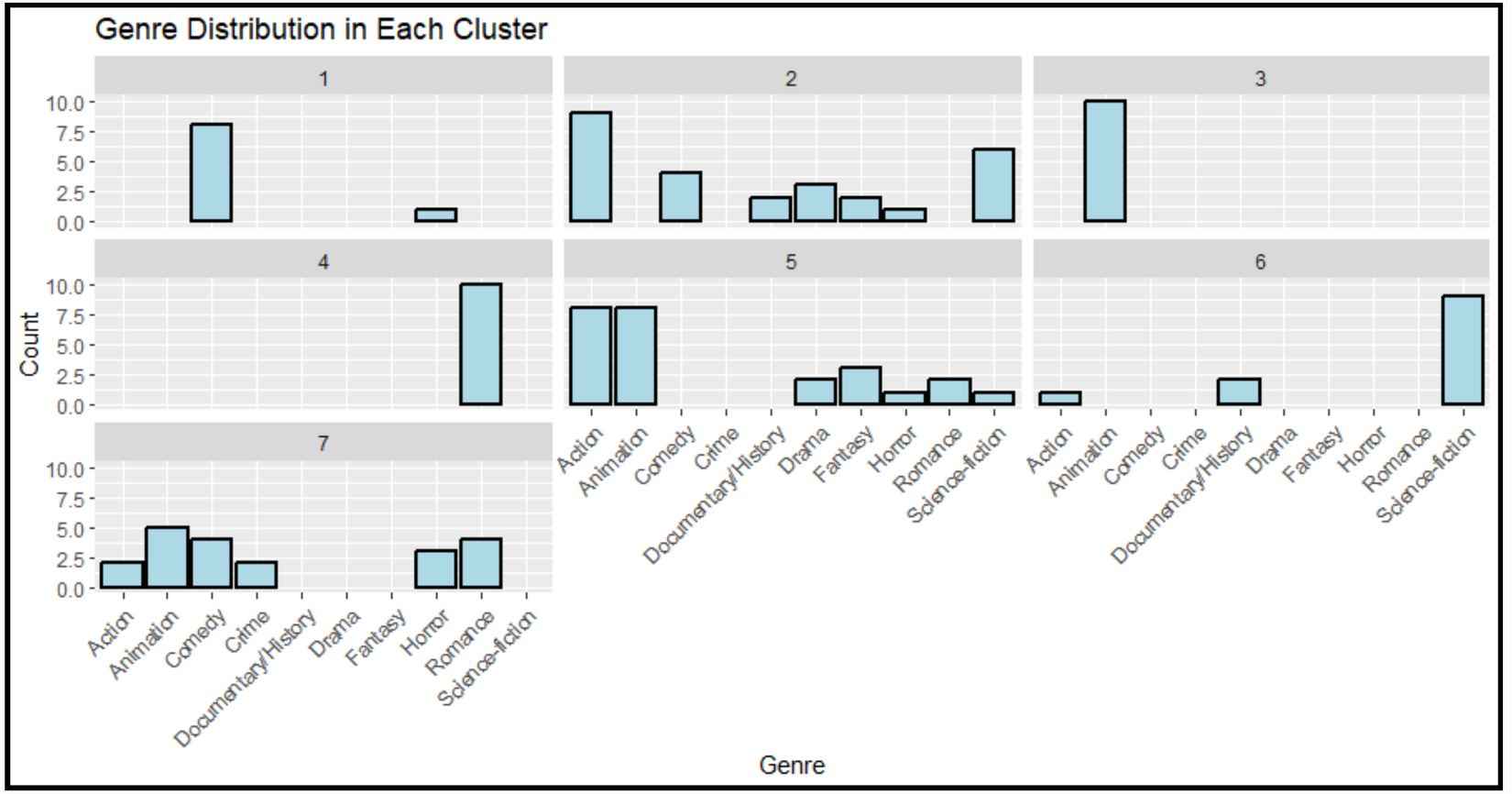
```r
# --- Data Preparation --- #
# add into numerical value
num_attr <- c("Age")

# add into categorical value
cat_attr <- c("Field", "Genre", "Factor")
df[cat_attr] <- lapply(df[cat_attr], as.factor)

# add into ordinal value
ord_attr <- c("Frequency")
df$Frequency <- factor(df$Frequency,
                       order = TRUE,
                       level = c("Less than 2 hours",
                                 "2 - 5 hours",
                                 "6 - 10 hours",
                                 "11 - 15 hours",
                                 "16 - 20 hours",
                                 "More than 20 hours"))

# put everything into a complete data set
process_dataset <- df %>% select(num_attr, ord_attr, cat_attr)
```

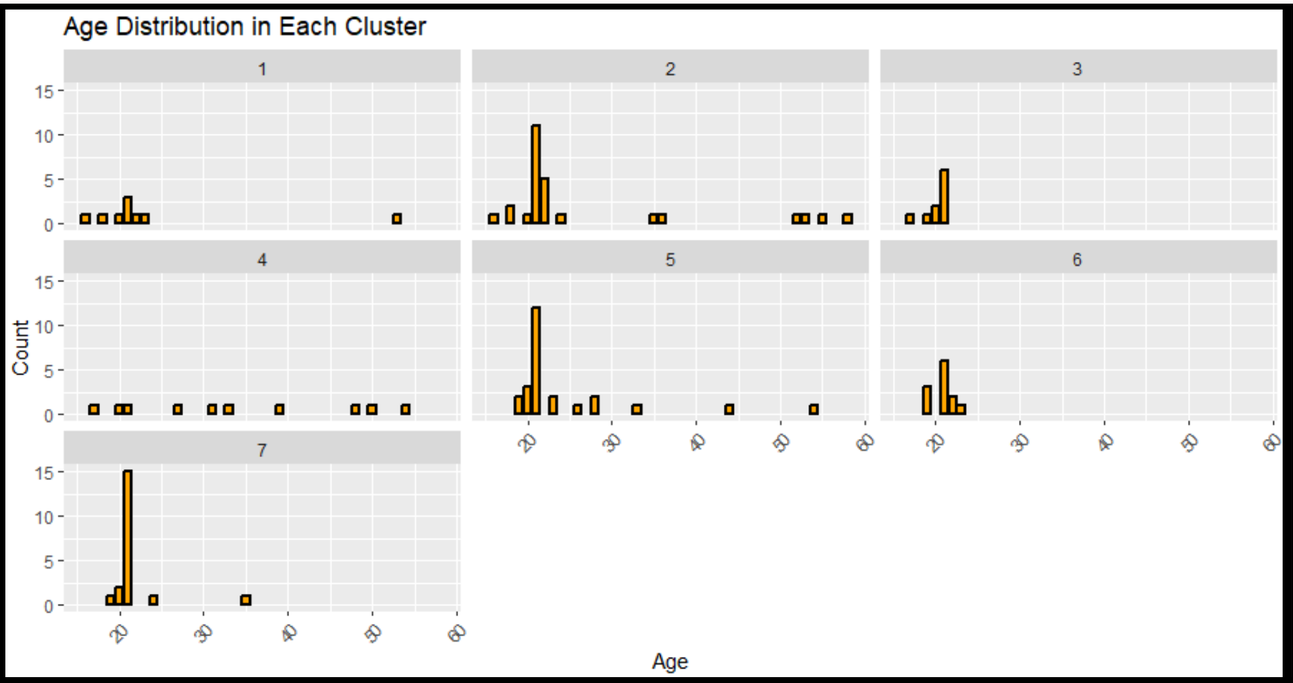
	Age	Frequency	Field	Genre	Factor
1	16	Less than 2 hours	Society	Comedy	The image and trailer of the movie
2	16	Less than 2 hours	Business	Comedy	The quality of the movie
3	17	6 - 10 hours	Society	Animation	Ratings of the movie
4	17	Less than 2 hours	Business	Romance	The cast and crew of the movie
5	18	2 - 5 hours	Society	Action	The quality of the movie
6	18	Less than 2 hours	Arts/Entertainment	Horror	The cast and crew of the movie

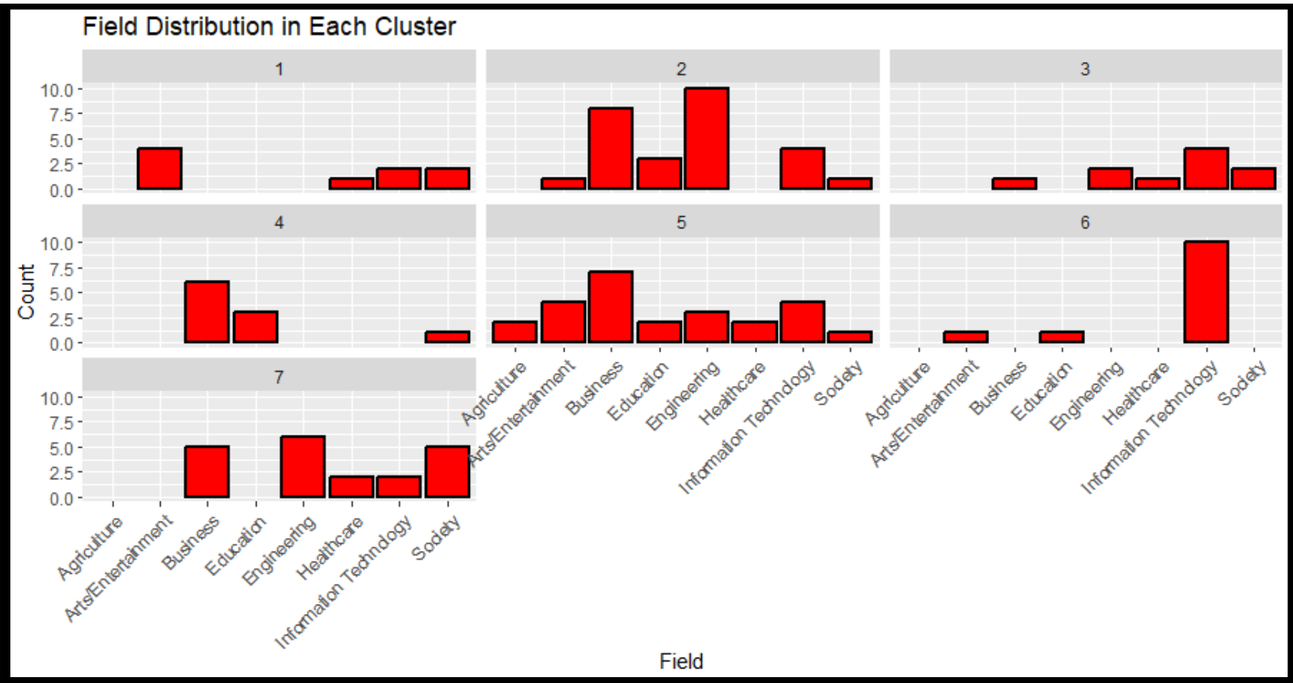
```r
          # --- Calculation --- #
# calculate Gower's distance
gower_dist <- daisy(process_dataset, metric="gower")

# hierarchical clustering, using ward.D method
gower_hcl <- hclust(gower_dist, method = "ward.D")

          # --- DENDROGRAM ---- #
# plot dendrogram
plot(gower_hcl, cex = 0.6)

# draw borders for the individual clusters
rect.hclust(gower_hcl, k = 7, border = 2:7)
```

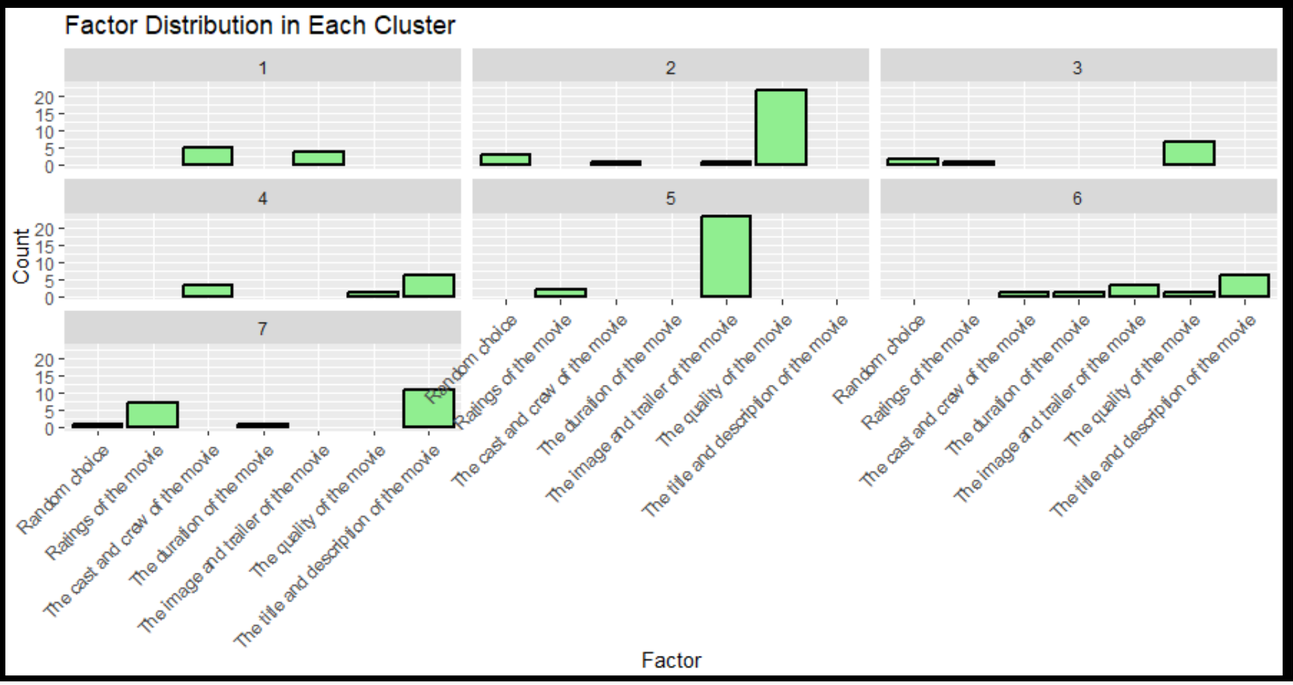


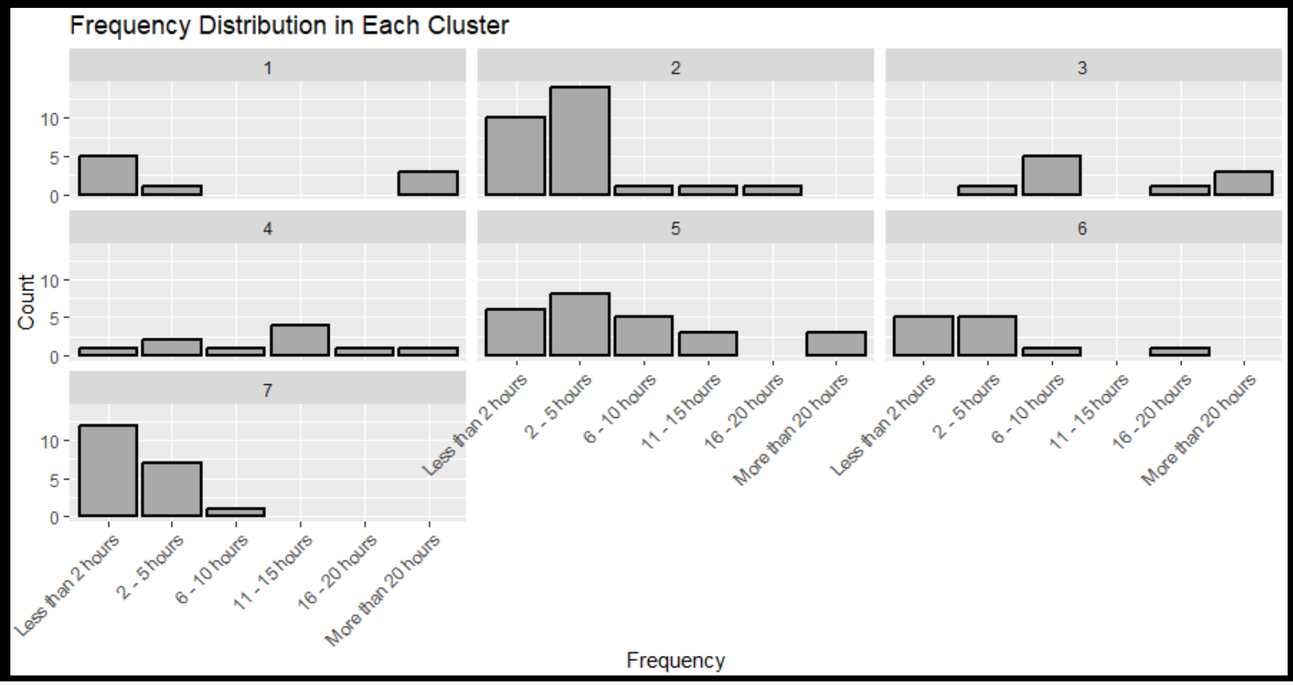
```r
          # --- HISTOGRAM --- #
# cut into k clusters
k <- 7
clusters <- cutree(gower_hcl, k)

# add the cluster assignments to the data frame
df$Cluster <- factor(clusters)

# histogram of Genre distribution in each cluster
ggplot(df, aes(x = Genre)) +
  geom_histogram(stat = "count", fill = "lightblue", color = "black", linewidth = 0.8) +
  facet_wrap(~ Cluster) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Genre Distribution in Each Cluster", x = "Genre", y = "Count")
```




Age Distribution in Each Cluster
1
2
3
4
5
6
7
Count
Age


Field Distribution in Each Cluster
1
2
3
4
5
6
7
Count
Agriculture
Arts/Entertainment
Business
Education
Engineering
Healthcare
Information Technology
Society
Field


Factor Distribution in Each Cluster
1
2
3
4
5
6
7
Count
Random choice
Ratings of the movie
The cast and crew of the movie
The duration of the movie
The image and trailer of the movie
The quality of the movie
The title and description of the movie
Factor


Frequency Distribution in Each Cluster
1
2
3
4
5
6
7
Count
Less than 2 hours
2 - 5 hours
6 - 10 hours
11 - 15 hours
16 - 20 hours
More than 20 hours
Frequency

05. Conclusion

Do we achieve the objectives?

Objective 1

Created a survey to collect and analyse data using R.

Objective 2

Used hierarchical clustering model to analyze, visualize the data.

Objective 3

Observed the clusters revealing similar patterns for the movie recommend system.

Cons of both methods

Dummy Variable

- **Curse of Dimensionality**: increase the dimensionality of the data
- **Unequal Distances**: assume equal distances between categories, which might not reflect the actual dissimilarities between them

Gower's Distance

Lack of documentation

Unexpected errors, hard-to-use functions

THANK YOU