VIETNAMESE-GERMAN UNIVERSITY
COMPUTER SCIENCE AND ENGINEERING

Compulsory Elective Subject: Data Analysis in High Dimensions

---

Group 4 Project Report

Hierarchical Cluster Analysis: Movie Genres Preferences

Instructor: Prof. Christina Andersson
Students: Vũ Hoàng Tuấn Anh (18812) Trần Kim Hoàn (18810)
Bá Nguyễn Quốc Anh (17965) Nguyễn Hoàng Hải Nam (17035)

BINH DUONG, NOVEMBER 2023

Member list & Workload

Our team has 4 members and we together decide to split the total workload into proper parts equally. Each member is responsible for the work as below:

No.	Full name	ID	Percentage of work
1	Vu Hoang Tuan Anh	18812	100%
2	Tran Kim Hoan	18810	100%
3	Ba Nguyen Quoc Anh	17965	100%
4	Nguyen Hoang Hai Nam	17035	100%

- **Vu Hoang Tuan Anh** (ID: 18812)
  - Collecting and preparing data
  - Applying hierarchical cluster model with dummy variables
  - Analysising results, plotting dendrogram and histogram, distribution diagram
  - Using Gap Statistics method to find the optimal number of clusters
  - Writing the report (4. Hierarchical clustering with Dummy variables)
- **Tran Kim Hoan** (ID: 18810)
  - Conducting the survey
  - Collecting and preparing data
  - Applying hierarchical cluster model with dummy variables
  - Analysising results, plotting dendrogram, histogram
  - Writing and finalizing the report (3. Data)
- **Ba Nguyen Quoc Anh** (ID: 17965)
  - Collecting data
  - Applying hierarchical clustering model with Gower's distance
  - Analysising results, plotting dendrogram, histogram
  - Writing the report (5. Hierarchical clustering with mixed-type variables + 6. Conclusion)
- **Nguyen Hoang Hai Nam** (ID: 17035)
  - Conducting the survey
  - Collecting data
  - Applying hierarchical cluster model with dummy variables
  - Writing the report (1. Project Objective + 2. Overview of Hierarchical clustering methodology)

Contents

1 Project Objective

We would like to group together users with similar viewing patterns in order to recommend similar content (genres). By asking different questions about age, genre, hobbies, we can collect data and conclude some insights.

2 Overview of Hierarchical clustering methodology

Hierarchical clustering is an algorithm that groups similar objects into clusters. The clusters are distinct from each other, and the objects within each cluster are broadly similar to each other.

Hierarchical clustering is an unsupervised learning technique. This means that a model does not have to be trained, and there is no need for a "target" variable.

There are two common categories of Hierarchical clustering, agglomerative and divisive.

- Agglomerative clustering starts in it individual clusters, then merges pair of clusters and continuing until all clusters have been merge in one huge clusters.

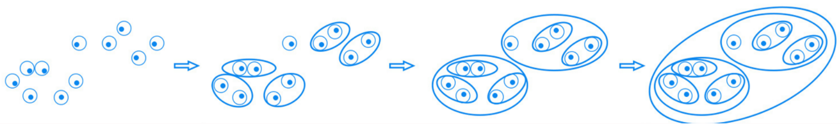


Figure 1: Agglomerative clustering

- Divisive clustering is the inverse approach of agglomerative clustering which starts in one cluster then splits into smaller cluster base on their difference.

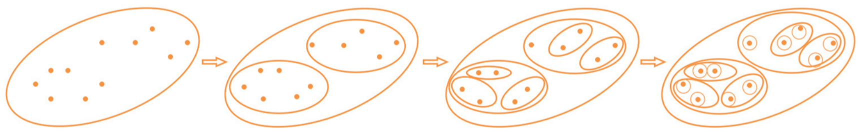
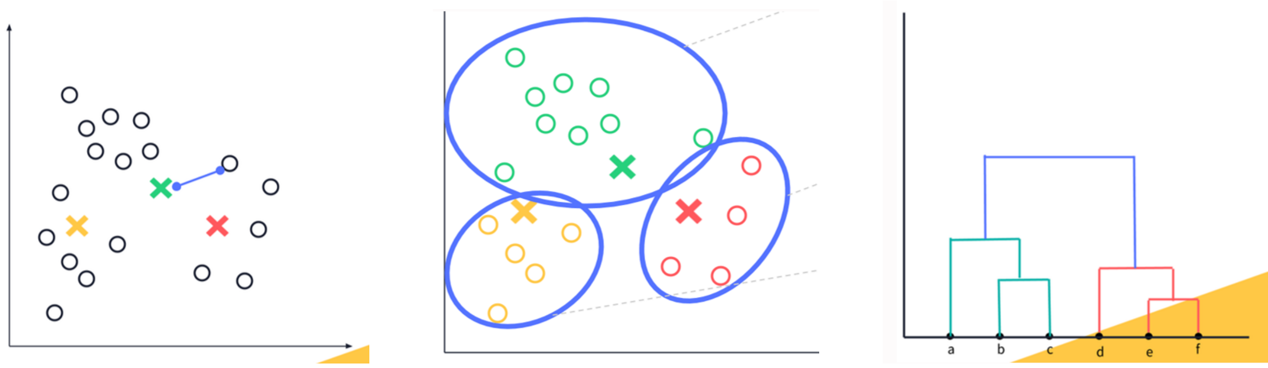


Figure 2: Divisive clustering

In our project, we only focus on agglomerative clustering.

The method of hierarchical clustering starts by treating each data point as a separate cluster and then iteratively combines the closest clusters until a stopping criterion is reached. The clusters are visually represented in a hierarchical tree called a dendrogram.



(a) Finding points distances (b) Grouping (c) Termination

Figure 3: Hierarchical clustering steps

1. **Finding points distances:** Calculate the distance between each point of observation. There are various mathematical algorithms to apply: "euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski"

2. **Grouping points (clustering)** Calculate the distance between each cluster. There are various mathematical algorithms to apply:"Ward", "single", "complete", "average", "mcquitty", "median", "centroid". After that, choose the optimal number of clusters.

3. **Termination** Plotting the final dendrogram and starting the analysis process based on attributes of each cluster.

3 Data

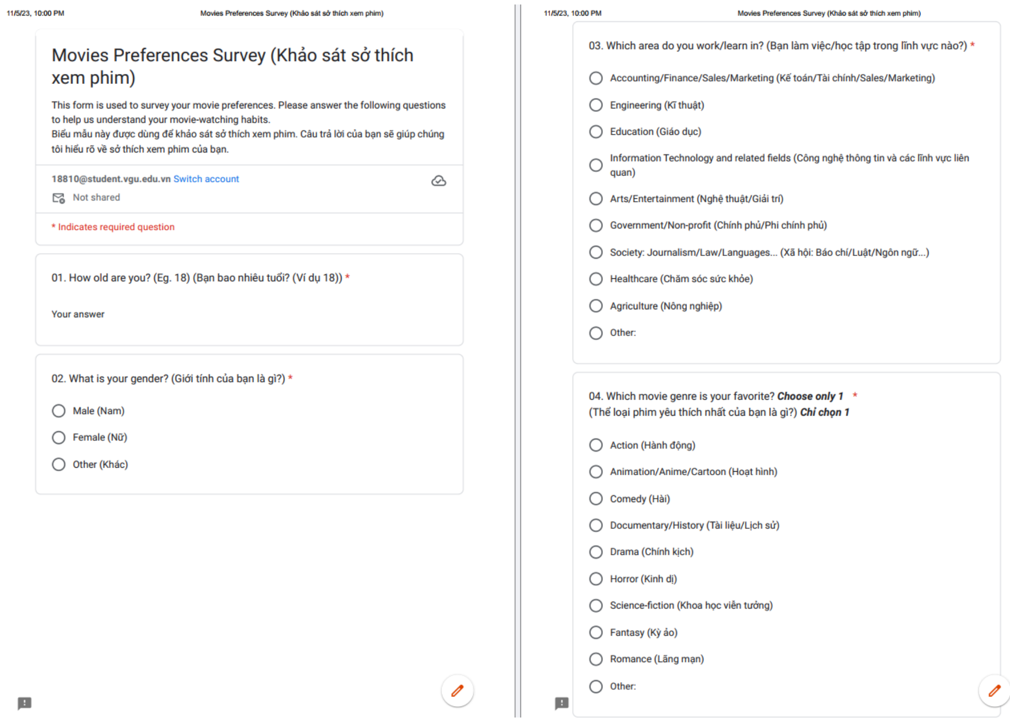
According to the objective of our project, we require data for our analysis. Our data is derived from our online survey conducted over a week. All collected data is accessible only to authorized research team members.

3.1 Attributes Identification and Selection

Initially, we brainstorm and identify key attributes which will be included in our analysis such as user demographics, user behavior and user preferences. After discussing their relevance and potential impact on the user's movie genre preferences, we eventually determine seven key attributes essential for our analysis: age, gender, working/learning area, preferred film genre, factor influencing genre choice, frequency of movie watching, source of film viewing.

3.2 Data Collection

- **Step 1: Create a survey**
  We create a survey titled 'Movie Genre Preferences' to collect data via Google Forms. In the form, we focus on seven key attributes that we determined before. (as discussed in Section 2.1).


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

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.

18810@student.vgu.edu.vn Switch account
Not shared

* 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)

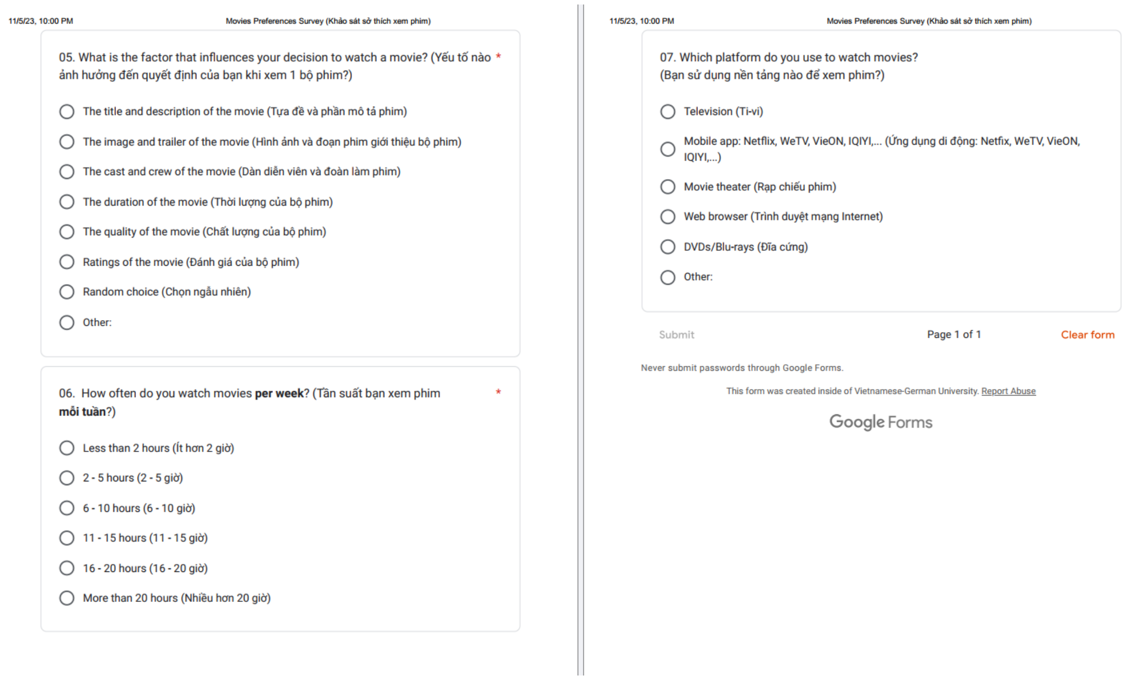
03. Which area do you work/learn in? (Bạn làm việc/học tập trong lĩnh vực nào?) *

- 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:

Figure 4: Movie genre preferences survey image (1)


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

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: Netflix, 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

Never submit passwords through Google Forms.

This form was created inside of Vietnamese-German University. Report Abuse

Google Forms

Figure 5: Movie genre preferences survey image (2)

- **Step 2: Publish survey and receive responses**

  After completing the form, we distribute it to individuals, specifically targeting those aged 16 and above. Subsequently, we received a total of 118 responses from the survey.

  Here is an image of the responses list:

	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 ı	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 m
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 m
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 m
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 m
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 m
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 m
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 pl
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 m
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 pl
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 m
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 m
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 m
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 m
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 m
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 m
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,

Figure 6: Image of the responses list

3.3 Data Preparation

This process is crucial in minimizing the potential for errors and inaccuracies that may arise during our data processing stage. By ensuring the data is accurately prepared and processed, we can enhance the reliability and validity of our analysis.

With a dataset of 118 responses, we come to the first crucial step for our data analysis: cleaning data.

- In terms of attributes, we convert them from question format to word/phrase format:

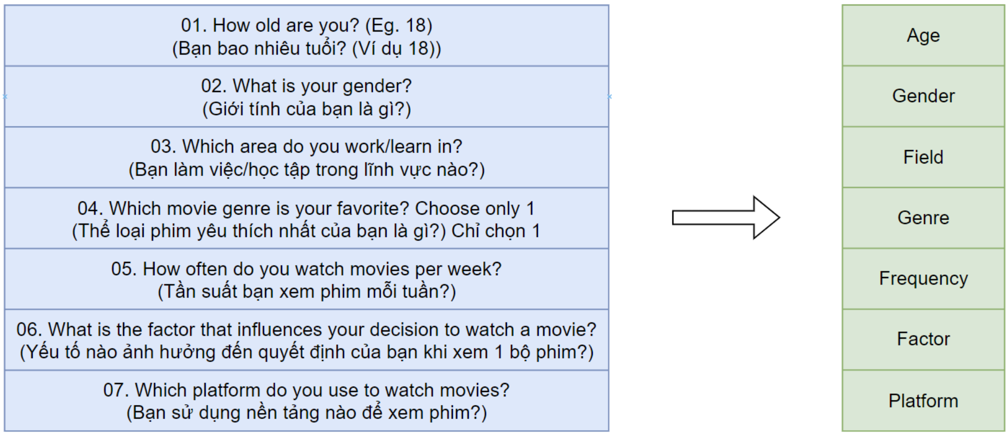


Figure 7: Image of attributes before and after

- We also substitute long words/phrases with shorter ones:

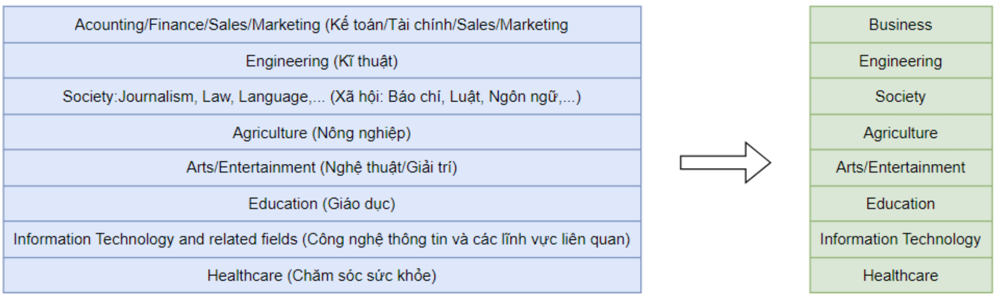
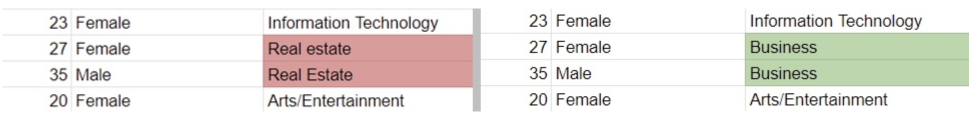


Figure 8: Image of the data before and after substitution

- Regarding working/learning area, we decide to combine "real estate" as a part of "business":



23	Female	Information Technology	23	Female	Information Technology
27	Female	Real estate	27	Female	Business
35	Male	Real Estate	35	Male	Business
20	Female	Arts/Entertainment	20	Female	Arts/Entertainment

Figure 9: Image of the data before and after combination

- We adjust responses that are not in the correct format into our anticipated format:

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

22	Male
16	Female
21	Male
16	Female

Figure 10: Image of the data before and after

- We remove responses that deviate from our expected range of values.

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

28	Female	Agriculture
21	Male	Education

Figure 11: Image of the data example before and after elimination

- The entry “housewife” does not correspond to a specific work or learning area. Therefore, we exclude it from our dataset.

52	Female	Housewife	Drama	11 - 15 hours	Random choice	Mobile app

Figure 12: Image of the entry "housewife"

After the data cleaning process, we obtain the final dataset. This dataset has the correct format, concise words/phrases, and values within an appropriate range for all attributes, making it suitable for our analysis.

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

Figure 13: Image of the final data after data cleaning process

At the next step, we apply the hierarchical clustering method with dummy variables and mixed variables with our final dataset.

4 Hierarchical clustering with Dummy variables

4.1 Problem

Regarding the objective of the project, users with similar viewing patterns need to be grouped together in a cluster, and in the previous section, we know that the dataset consists of both numerical and categorical data. Especially, it is impossible to calculate the distance between any 2 entries of the data frame by applying mathematical algorithms (Euclidean, Manhattan, Canberra, etc) directly since all variables are not numerical. To handle this situation, an encode is needed to convert variables from nominal to numerical type. (The conversion from ordinal variables to numerical variables is redundant since ordinal variables could be labeled with levels and then be scaled to behavior as numerical variables). That encoding is called ***Dummy variables***, and this section will discuss the implementation of dummy variables for hierarchical clustering.

4.2 Dummy variables

- **Definition:** A Dummy (or Indicator) variable is an artificial variable created to represent a categorical variable with two or more distinct categories or levels.
- **Usage:** With dummy variables, all nominal variables could be converted to numerical. Hence, it is possible to apply mathematical algorithms to start the analysis process.
- **Example:** A nominal variable named "Genre" always takes one of 3 values "Action", "Comedy", and "Fantasy". After converting it into the dummy variable:
  - If the value of the variable is "Action", the indicator for "Genre.Action" attribute will be 1, and other attributes are 0's.
  - If the value of the variable is "Comedy", the indicator for "Genre.Comedy" attribute will be 1, and other attributes are 0's.
  - If the value of the variable is "Fantasy", the indicators for all attributes are 0's.

Genre
Action
Comedy
Fantasy

(a) Original "Genre" variable

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

(b) "Genre" Dummy variable

Table 1: Using Dummy variables

→ A variable which has a total of n different values will have new $n - 1$ attributes after converting to a dummy variable, and the last value has the indicators for all attributes are 0's.

4.3 Applying the Hierarchical clustering model

- In this section, R programming language is used as the main language to implement the hierarchical clustering model.
- Dependencies (List of packages used):

Package name	Description	Version
base	The R Base Package	$\geq 4.3.2$
cluster	Finding Groups in Data	$\geq 2.1.4$
factoextra	clustering visualization	$\geq 1.0.7$
gplots	plotting data, dendrogram	$\geq 3.1.3$
ggplot2	draw distribution graph	$\geq 3.4.4$

(Source code link: https://github.com/vhtuananh020402/Group4_Data_analysis/blob/main/dummy_canberra_ward.r)

To start applying the hierarchical clustering model, first, the dataset is imported and stored as a data frame in R. The function ***na.omit()*** is necessary for omitting the missing value of the imported data frame.

```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)
```

Then, we standardize (scale) the "Age" (numerical) variables. Since the "Frequency" variable is ordinal, it first needs to be ordered and labeled with levels, then converted to a numerical variable using ***as.numeric()*** function. After that, the "Frequency" (numerical) variable is scaled.

```r
# 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)
```

The function ***model.matrix()*** is used to create a data frame matrix consisting of all numerical variables (all nominal variables "Genre", "Field", "Factor", "Gender", "Platform" are converted to dummy variables)

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

Here we use Canberra algorithms to calculate the points distance and Ward's method to calculate the clusters distance. Then we can visualize the hierarchical clusters by using the function ***plot()*** to plot a dendrogram.

```
# 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

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

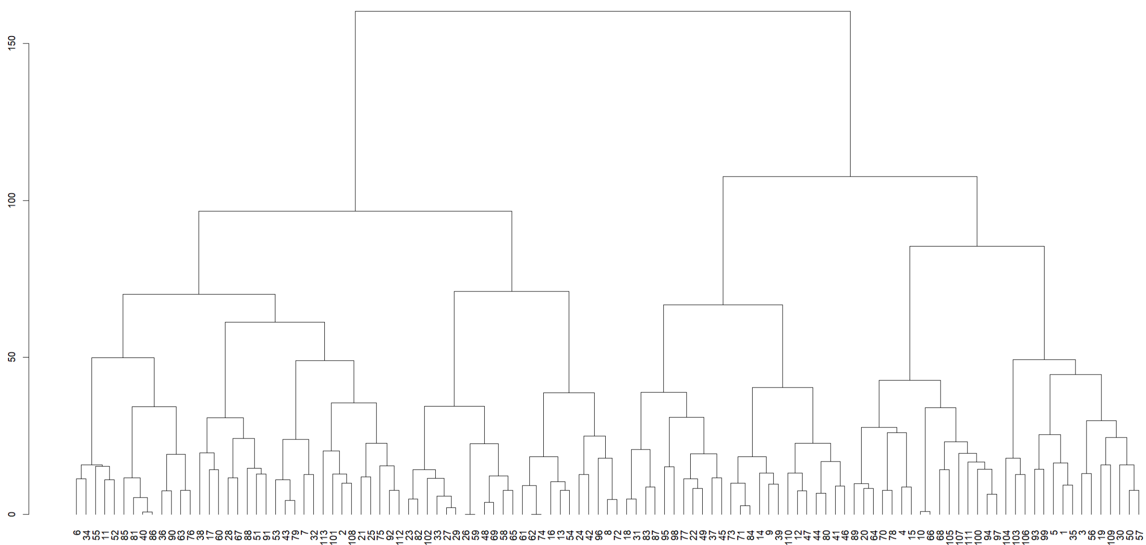


Figure 14: Dendrogram visualization

4.4 Finding the optimal number of clusters

After dendrogram visualization, as can be clearly seen that there are many possible numbers of clusters to choose from for analysis. Each different number of clusters will bring different outcomes, thus we need an exact optimal number of clusters, and there are some methods that help us to find it, **Gap Statistics** is one of the most well-known and easily understandable methods to find the optimal number of clusters, so this section will focus on it.

4.4.1 Gap statistics 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 code implementation:

```r
# ======= Gap statistic method to find the optimal number of cluster  =======
# Calculate Within-Cluster Dispersion (WCD) for the original data
wss <- sum(hc$height)

# Generate Random Data for Comparison
set.seed(123)    # Set seed for reproducibility
B <- 100         # Number of random datasets
random_datasets <- lapply(1:B, function(i) matrix(runif(length(df_dummy)), ncol =
ncol(df_dummy)))

# Cluster the Random Data
random_hcs <- lapply(random_datasets, function(random_data) {
  dist_matrix <- dist(random_data, method = dist_method[4])
  hclust(dist_matrix, method = hc_method[1])
})

# Calculate Within-Cluster Dispersion for Random Data
wss_random <- sapply(random_hcs, function(random_hc) sum(random_hc$height))

# Calculate Gap Statistic
gap <- (log(wss_random) - log(wss)) + mean(log(wss_random) - log(wss))

# Determine the Optimal Number of Clusters
num_clusters <- 1 : 10 # Desired number of cluster domain
```

```
gap <- gap[1:length(num_clusters)]
x_labels <- seq(1, length(num_clusters))

plot(num_clusters, gap, xlab = "Number of Clusters", ylab = "Gap Statistic",
     main = "Gap Statistic Plot", type = "b")

# Customize x-axis labels
axis(1, at = x_labels, labels = num_clusters)
abline(v = 8, col = "red", lty = 2)
```

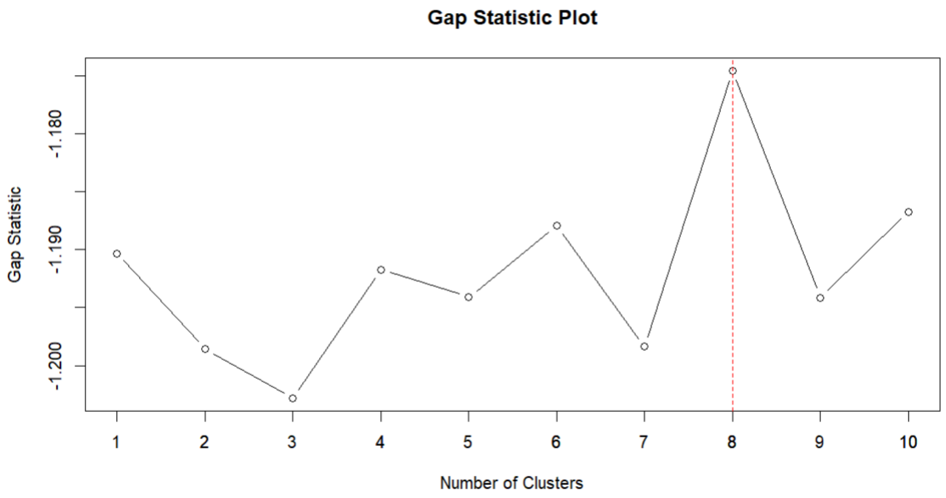


Figure 15: Gap Statistics diagram

→ From the above Gap Statistics diagram (range 1 : 10), the highest point which indicates the optimal number of clusters for the dendrogram is 8. Hence, we chose the number 8 for the number of clusters for the dendrogram.

4.5 Result

4.5.1 Dendrogram

```
# Draw the rectangle around each cluster in k clusters
k <- 8
rect.hclust(hc, k, border = 2:8)
```

Here is the final dendrogram with 8 clusters, each cluster has a rectangle around it. The observations in each cluster have a close relationship, which will be discussed right after this dendrogram.

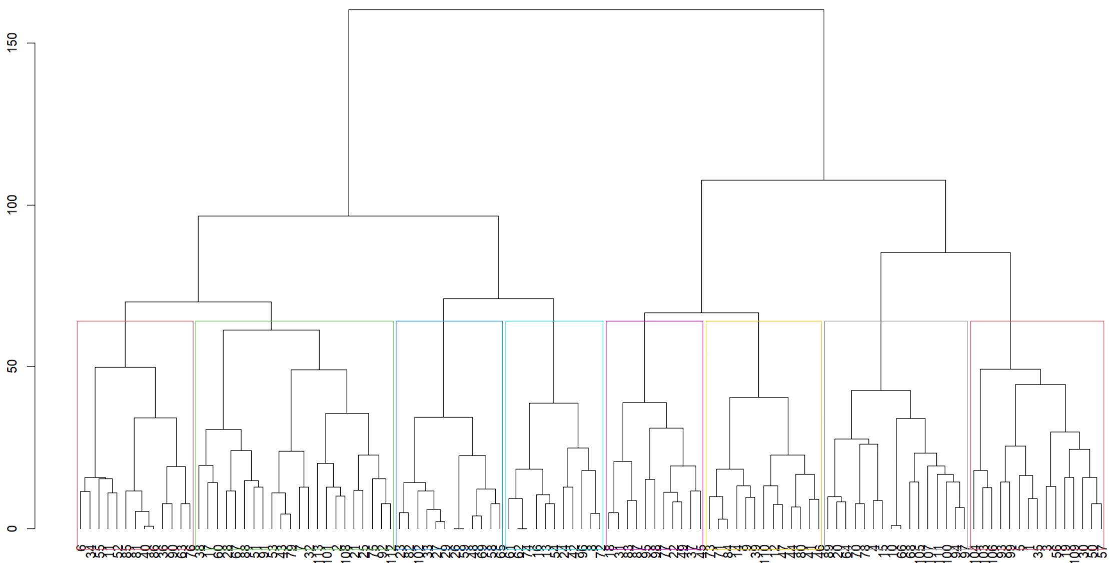


Figure 16: Age Distribution in each cluster

4.5.2 Histogram and observing patterns

"Age", "Genre", "Field", "Frequency", and "Factor" are considered the main attributes of the movie genre recommender system, then we observe the patterns of those attributes by histograms.

R code to plot the histograms for each cluster:

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

# Create a histogram of the 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")

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

```r

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

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

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

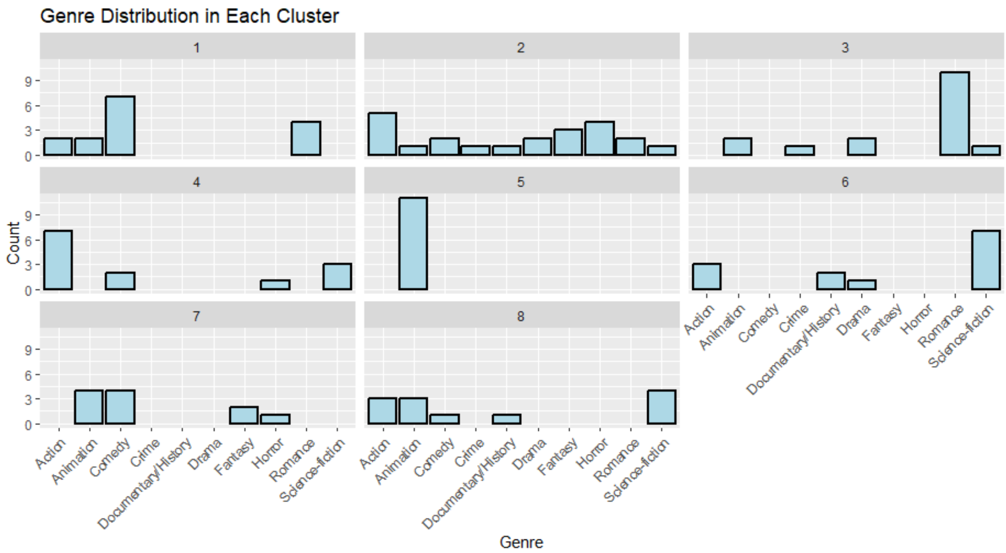


Figure 17: Genre Distribution in each cluster

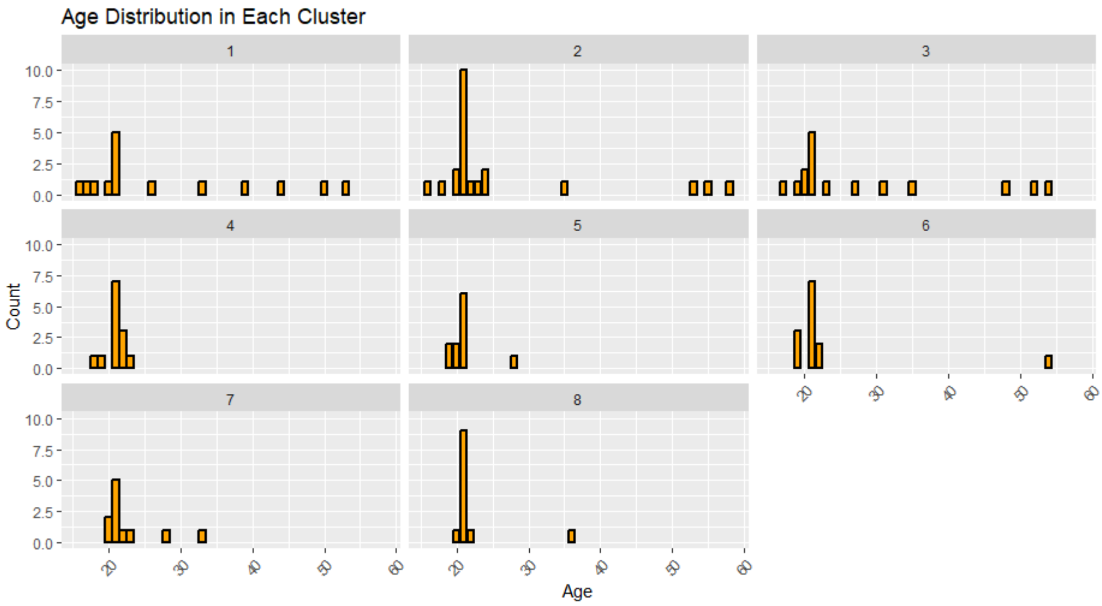


Figure 18: Age Distribution in each cluster

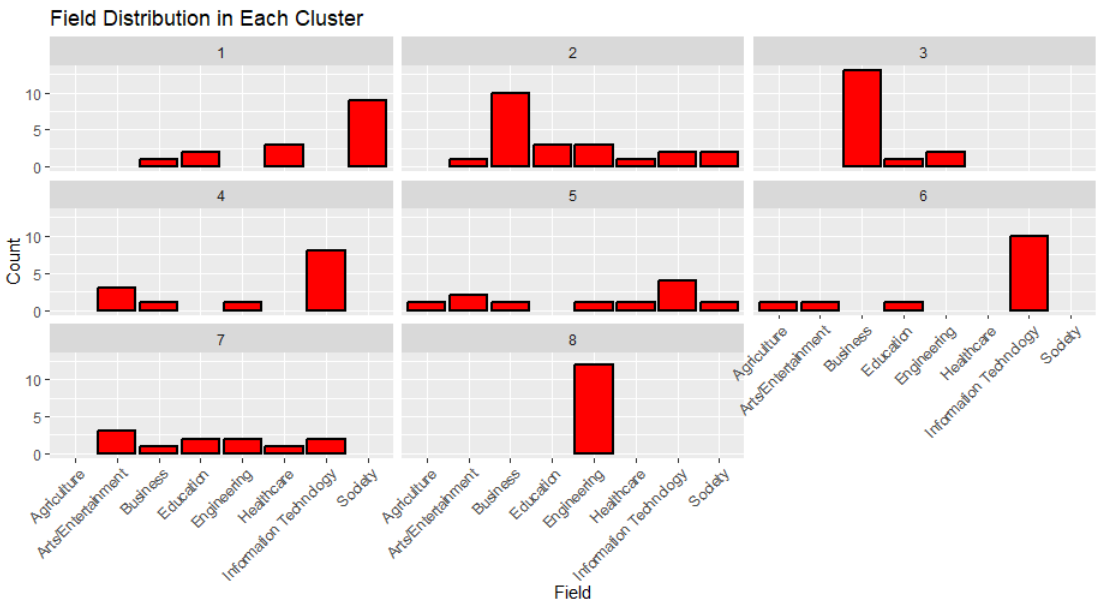


Figure 19: Field Distribution in each cluster

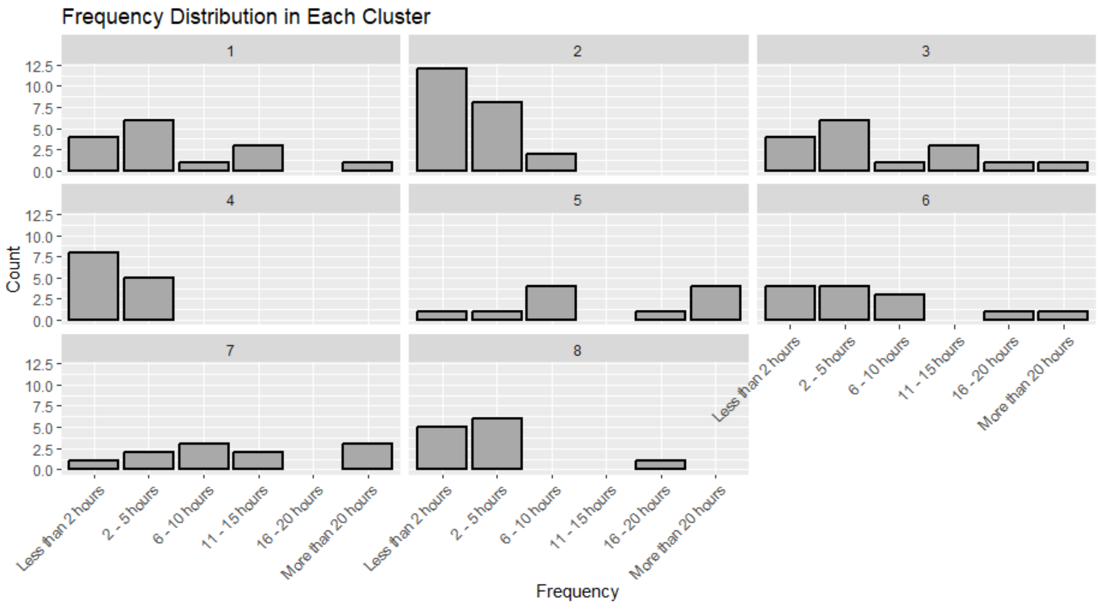


Figure 20: Frequency Distribution in each cluster

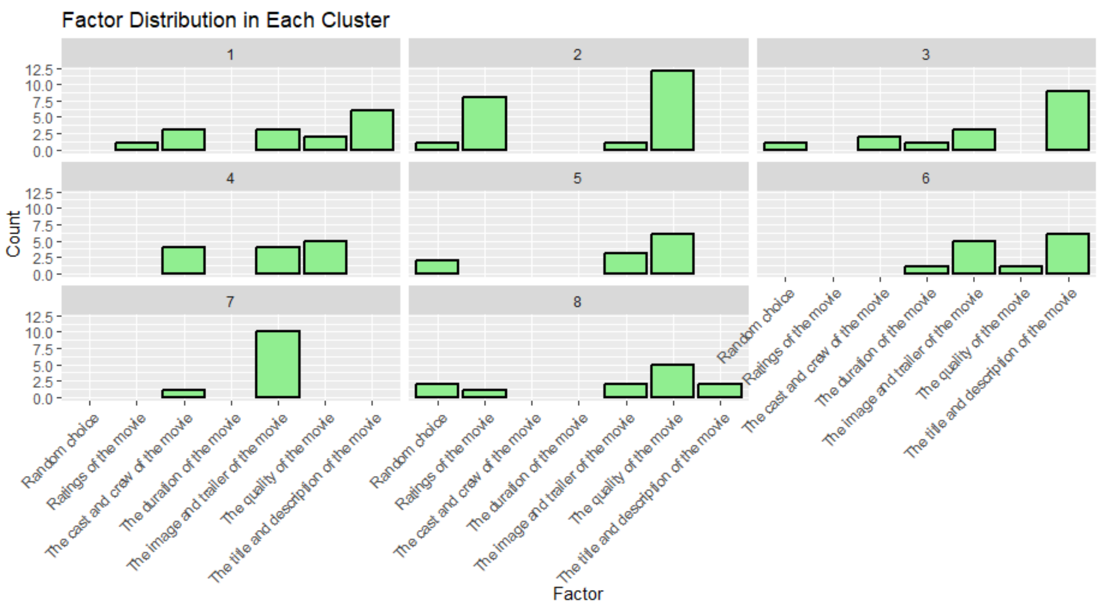


Figure 21: Factor Distribution in each cluster

From the histograms, here are conclusions for each cluster:

- **<u>Cluster 1:</u> Comedy** is the most favorite genre followed by **Romance**, with the majority of viewers being 21 years old and a small number ranged from 22 to 55 years old. Watching time is from 2 to 5 hours per week. Society field is the highest choice here. “Title/Description” is the most popular reason for choosing and watching.

- Cluster 2: **Action**, **Horror**, and **Fantasy** are top-viewed genres (Romance and Drama are also considered). The viewers are mostly under 25 years old. The average watching time falls between “Less than 2 hours” and “2 - 5 hours” per week. The number is largely come from Business and evenly distributed in the Education and Engineering fields. “Ratings” and “Quality” are the two most important factors for picking movies.

- Cluster 3: **Romance** is the top-most choice of movie genre in this cluster. The viewers are divided into 2 main age groups: around 25 and above 45 years old. Watching times are “2 - 5 hours” and “11 - 15 hours”. Business is the highest-picked field. “Title/Descriptions” is the most important factor.

- Cluster 4: **Action** and Science-fiction are the most favorite genre, with the majority of viewers being under 21 years old. Watching time is “Less than 5 hours” per week. IT field is the highest choice here, followed by Arts/Entertainment. “Cast/Crew”, “Image/Trailer”, and “Quality” are the most popular reasons for choosing and watching.

- Cluster 5: It is clearly seen that **Animation** is the only choice of movie genre in this cluster. The viewers are mostly 19 - 21 years old, and a small number is from 26 - 27 years old. The average watching time is pretty high and falls between “6 - 11 hours” and “More than 20 hours” per week. The number is largely come from IT and evenly distributed in the Agriculture, Arts/Entertainment, Business, Engineering, Healthcare, and Society fields. “Image/Trailer” and "Quality" are the two most important factors for picking movies.

- Cluster 6: **Action** and **Science-fiction** are the top-most choices of movie genres. Viewers are mostly around 20 and a small one is above 50 years old. Watching time is total less than 10 hours per week. IT is the highest-picked field. “Image/Trailer” and “Title/Descriptions” are the most important factor.

- Cluster 7: **Animation** along with **Comedy** is the most favorite genre followed by **Fantasy** and **Horror**, with the majority of viewers being under 35 years old. Watching time varies from less than 15 hours and more than 20 hours per week. The agriculture field is the highest choice here, and there exists evenly distribution of Education, Engineering, and IT fields. “Image/Trailer” is the most popular reason for choosing and watching.

- Cluster 8: **Action**, **Animation**, and **Science-fiction** are top-viewed genres. Viewers are mostly under 25 years old. The average watching time falls between “Less than 2 hours” and “2 - 5 hours” per week. The number is all come from Engineering fields. “Quality” and “Title/Description” are the two most important factors for picking movies.

4.6 Advantages and Disadvantages: Dummy variables

4.6.1 Advantages

- **Nominal Data Handling:** Hierarchical clustering algorithms typically work with numerical distances or similarities between data points. Dummy variables allow us to represent nominal variables numerically, enabling their incorporation into the clustering process.

- **Information Preservation:** Dummy variables can help preserve information about the nature of the data. By creating separate binary variables for different categories, we maintain the distinctions between nominal variables during clustering.

4.6.2 Disadvantages

- **Unequal Distances**: Dummy variables assume equal distances between points, which might not reflect the actual dissimilarities between them.

- **Increased Noise:** If the nominal variable has a large number of attributes with limited observations in each attribute, the dummy variables may introduce noise into the clustering process, leading to less reliable results.

5 Hierarchical clustering with mixed-type variables

5.1 Problem

One important aspect of processing information is to deal with different types of data, because not every attribute is of the same nature.

Given an example: A survey is conducted about a person's daily life. The considered factors are: Age, Gender, Job, and how Satisfactory they are. A quick observation shows that these values are not the same type. Age is measured in integer, which is a numerical value; Gender can be used as a binary value, with 0 (false) being female, and 1 (true) being male; Job is shown with many options (engineer, businessman, storekeeper, homemaker, ...), therefore it is a nominal value; and a scale of 1 to 10 to rate the person's satisfaction, which is an ordinal value. This is just some common types of data that we usually deal with, so knowing how to process mixed data can benefit us with a more realistic model.

This section will discuss Gower's Distance, which is one of the methods to deal with this problem.

5.2 Gower's Distance

Gower's Distance is a method for computing distances between two data points. The strength of this method comes from the fact that it is usable for other types of data beside numerical, which is more flexible than the common methods (Euclidean distance, Manhattan distance, ...). Another advantage is that Gower's Distance scales the ranges of data into between 0 and 1, with addition of allowing a user-defined weighting scheme. But for this project, an unweighted model is constructed.

The basic calculation of Gower's Distance is as follow. Data will be separated into two types: numerical and non-numerical.

With numerical, we can compute the data by the formular:

|Difference| / Range

- Difference = Data[i] - Data[i+1] ;
- Range is the difference between the maximum and the minimum data points.

With non-numerical values, we can compute by compare the data points. If they are identical, the distance will be 0, if not then the distance will be 1.

There are some packages that uses or have the option for Gower's calculation, and in this project "cluster" package is used, which contains function "daisy" that allows Gower as an option.

5.3 Code Explaination

(Source code link: https://github.com/vhtuananh020402/Group4_Data_analysis/blob/main/completed_gower_ward.r)

First, libraries and data frame will be imported. The function na.omit(df) is necessary for omitting missing values in the data frame.

```
library(factoextra)   # clustering visualization
library(ggplot2)      # draw distribution graph

# read data frame
df <- read.csv("data/clean_data_v2.csv")
# remove missing values in data frame
df <- na.omit(df)
```

Then, data will be processed by splitting into 3 types: numerical, nominal, and ordinal. Nominal is defined by using function lapply(df[nom_attr], as.factor) to change value type into factor. Ordinal is defined by using function factor(), which has the option order = TRUE, and level is from lowest to highest. Finally, process_dataset will contain the complete dataset for analysing steps.

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

# add into nominal value
nom_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)

head(process_dataset)
```

Function daisy() will calculate the dissimilarity matrix, with process_dataset as input, and gower is used as the calculation metric. After that, the hierarchical clustering model is built using hclust() function. Clustering method ward.D is chosen because it brings the best result when plotting dendrogram.

```
    # --- 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)
```

Histograms are used to present the distribution of each attribute in each cluster. “k” represents the number of clusters that we want. The decision k = 7 is made according to the previous method of building hierarchical clustering using Dummy Variable. For some unknown reasons, the functions that find the optimal number of clusters is unusable in this section.

```
    # --- 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")

# histogram of Field distribution in each cluster
ggplot(df, aes(x = Field)) +
```

```
    geom_histogram(stat = "count", fill = "red", color = "black", linewidth = 0.8) +
    facet_wrap(~ Cluster) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Field Distribution in Each Cluster", x = "Field", y = "Count")

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

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

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

5.4 Results

5.4.1 Dendrogram

The dendrogram shows the relations between all of the observations, with each cluster is shown by colored lines. With many tries and many different methods, we decided to keep this result because this is more evenly distributed model than other versions.

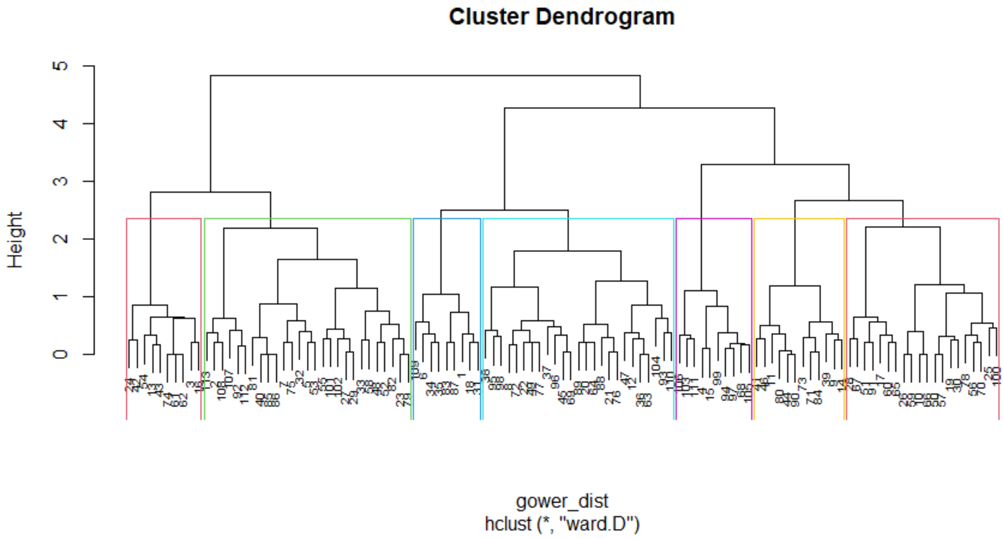


Figure 22: Result visualized by Dendrogram

5.4.2 Histograms and observing patterns

We only obverse the histograms of the following attributes: Age, Genre, Field, Frequency and Factor. Gender and Platform are useful for unifying the dataset, but those are not considered to be the attribute that contributing to the recommended system in real-world practices. Therefore, we omit them when graphing the histograms.

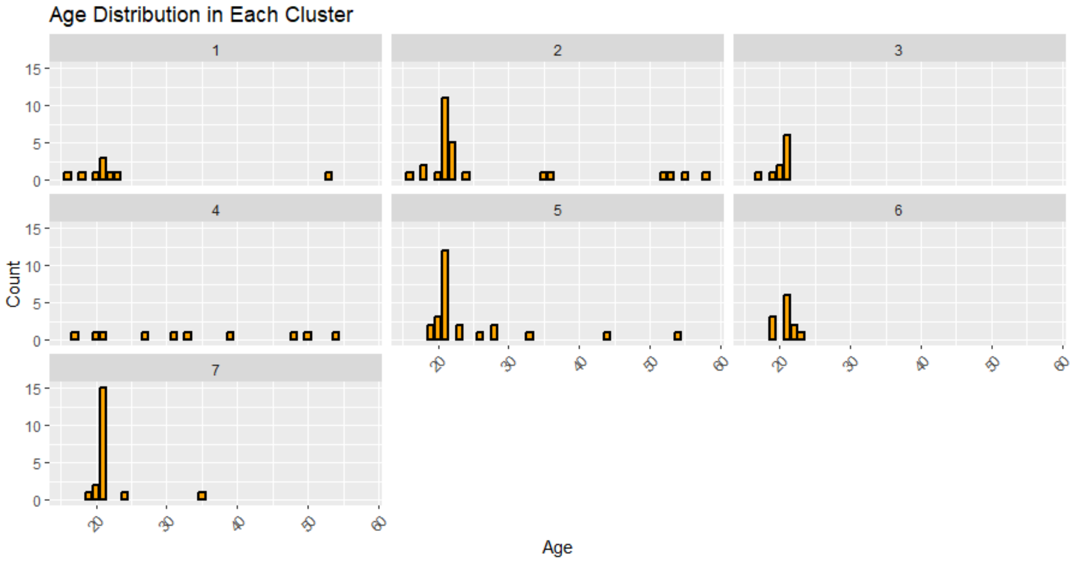


Figure 23: Age distribution in each cluster visualized by Histogram

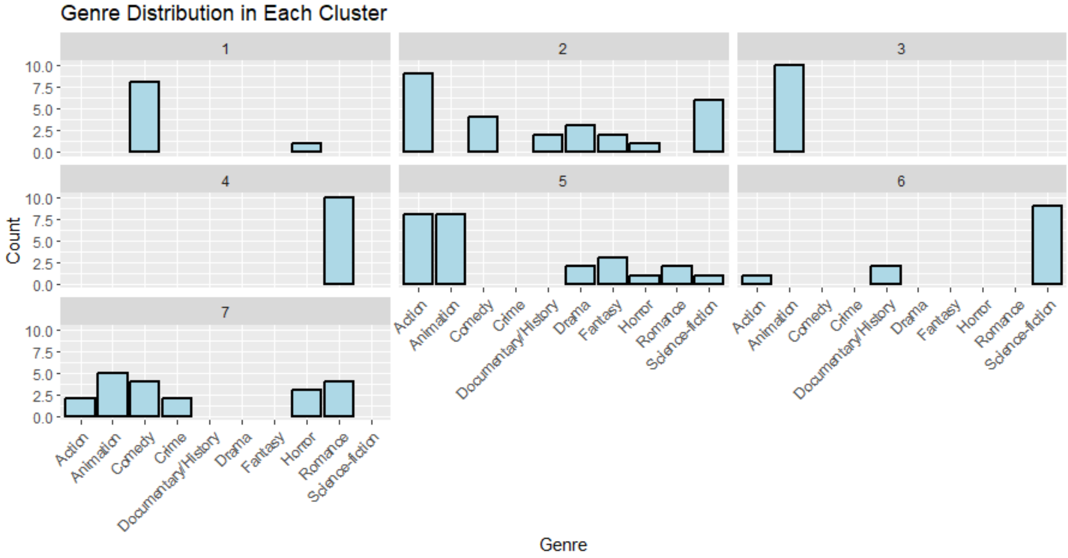


Figure 24: Genre distribution in each cluster visualized by Histogram

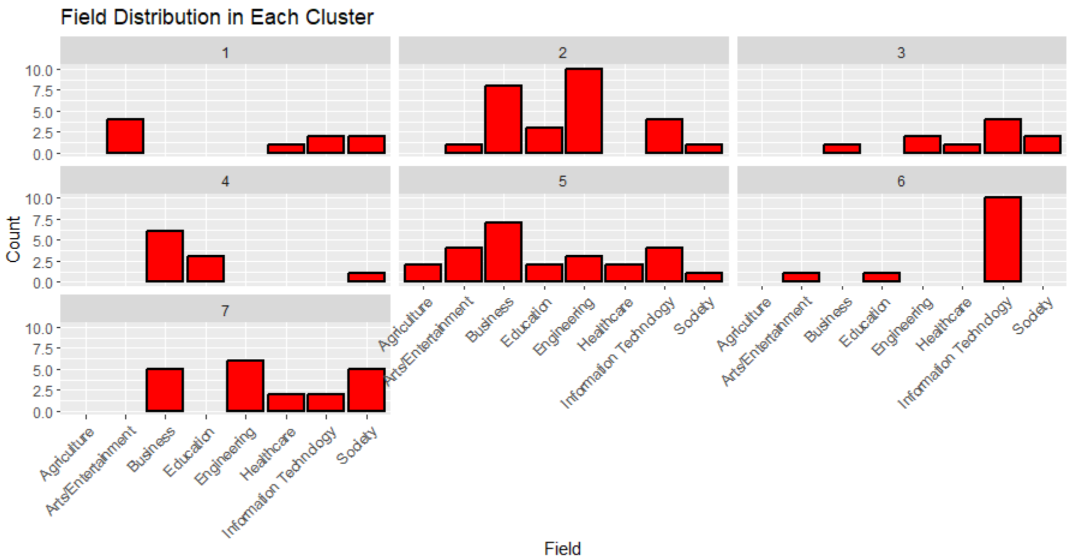


Figure 25: Field distribution in each cluster visualized by Histogram

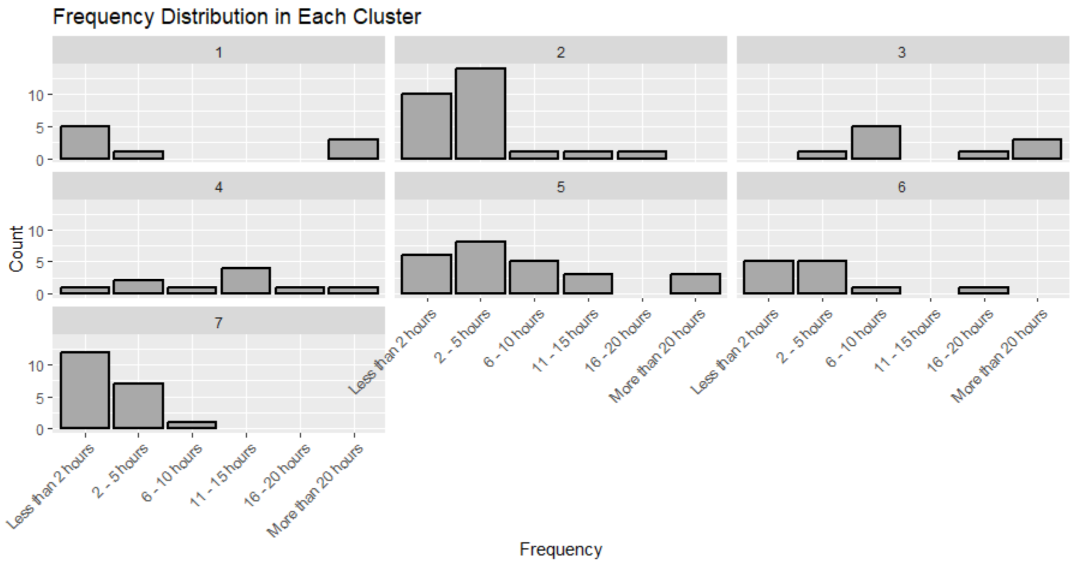


Figure 26: Frequency distribution in each cluster visualized by Histogram

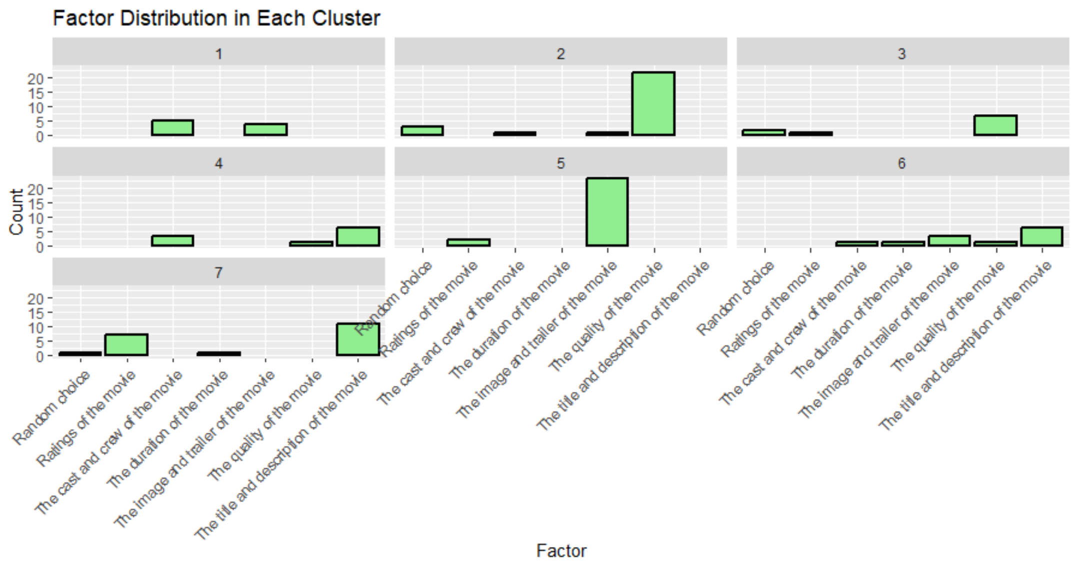


Figure 27: Factor distribution in each cluster visualized by Histogram

The followings are concluded from the histogram for each cluster:

Cluster 1: Mostly from under 25 years old. Comedy is the most watched genre, with watching time falls between "Less than 2 hours" and "Over 20 hours". Art/Entertainment field being the highest choice here, with IT and Society being the lesser. "Cast/Crew" and "Image/Trailer" are the reasons for choosing and watching.

Cluster 2: The majority are 21 years old, with Action and Sci-fi are the most viewed. Comedy, Documentary/History, Drama, Fantasy and Horror are also considered. Watching time is from 2 to 5 hours. They are largely come from Engineering and Business. "Quality" is the most important factor for picking movies.

Cluster 3: The majority are 21 years old, with Animation being the only choice for genre. "6 - 10 hours" and "more than 20 hours" are the standout choices. Engineering, IT and Society are evenly distributed fields, with IT being the highest choice. "Quality" is the most important factor.

Cluster 4: Ages are evenly distributed across the range. Romance is the only viewed genre here. Watching time is variety, with "11 – 15 hours" being the highest choice. Business is the most picked field here, and they consider "Cast/Crew" and "Title/Description" when choosing movies.

Cluster 5: Mostly 21 years old, with Action and Animation are the highest viewed genre here, also with somewhat evenly distributed between Drama, Fantasy, Romance. Watching time is largely from "Less than 2 hours" to "6 – 10 hours" range. Business is the most choice for field, but the occupation is distributed across the range. "Image/Trailer" is the most important factor.

Cluster 6: Around 20 years of age. Science-fiction is the most viewed genre, with small percentage comes from Action and Documentary/History. Mostly from "Less than 2 hours" to "2 - 5 hours" range of watching time. IT field is the highest choice, with "Image/Trailer" and "Title/Description" are the most important factor.

Cluster 7: The majority are 21 to 22 years old. Action, Animation, Comedy, Crime, Horror and Romance shared the genre distribution without much differences. Watching periods mostly come from "Less than 2 hours" to "2 – 5 hours" range. Business, Engineering and Society are the most choices for field of occupation, and "ratings" as well as "title/description" are the most important factors.

5.5 Shortcomings

With the lack of documentation, using libraries that implement Gower's Distance is complicated and troublesome, as it leads to unexpected errors. One example is that functions to find the optimal number of clusters cannot be applied, whether it is Elbow, Silhouette or Gap Statistic method. Because of the limited time frame of this project, the errors cannot be solved yet. Therefore, it is crucial to research thoroughly beforehand if using the existing libraries for this method.

6 Conclusion

6.1 What do we achieve objectives

From conducting the survey, preparing data, analyzing input, to visualize attributes by histograms, our team can identify the relation between viewing patterns(age, film genre, working field, frequency, factor) and we can recommend contents base on that insight. For that reason, our team have achieved our project goals and objectives.

6.2 Future Work

Refining clustering algorithms, adding more features, or expanding the recommendation system's capabilities are admirable improvements for this project. Other potential areas can be building the movie recommendation web app, testing our recommendation, accuracy; as well as implementing other methods to analyse data such as non-hierarchical clustering, comparing the result between two methods.

6.3 Project Repository on Github

All source codes and data files of this project can be found in the project repository on Github. Here is the link to this project repository on Github: https://github.com/vhtuananh020402/Group4_Data_analysis

References

[1] D. McCaffrey J. Example of calculating the gower distance. 2020. URL: https://jamesmccaffrey.wordpress.com/2020/04/21/example-of-calculating-the-gower-distance/.

[2] Alboukadel Kassambara. Determining the optimal number of clusters: 3 must know methods. URL: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/.

[3] D'Orazio M. Distances with mixed-type variables, some modified gower's coefficients. 2021. URL: https://arxiv.org/abs/2101.02481.

[4] RDocumentation. daisy: Dissimilarity matrix calculation. URL: https://www.rdocumentation.org/packages/cluster/versions/2.1.4/topics/daisy.

[5] Smita Skrivanek. The use of dummy variables in regression analysis. 2009. URL: https://www.moresteam.com/resources/whitepapers/dummy-variables.