Hierarchical clustering in gaming habit of Vietnamese - German University students

Nguyen Duy Phu Quang - 15890
15890@student.vgu.edu.vn

Le Duc Minh - 16669
16669@student.vgu.edu.vn

Le Thi Ngoc Diep - 16090
16090@student.vgu.edu.vn

Truong Nguyen Thien An - 16772
16772@student.vgu.edu.vn

Contents

List of Figures

Chapter 1: Introduction

1.1 About gaming habit

The online video game industry has grown exponentially in the past few decades. They have become an integral part of the lives of people all around the world. The question of why people play digital games has long been a point of both scholarly and public interest. Typically, the research has focused on personal motives for gameplay. There have been topics [1, 2] that categorize players into "achievers", "explorers", "socialisers", and domination-oriented "killers" based on their gameplay preferences is the first, or at least the earliest influential, example of digital game player taxonomies. The online game developer's goal is to profit from players downloading or playing the game. For example, in Europe, the German gaming market is in the top two only behind Russia in terms of size and revenue. Estimates are that there are approximately 44.3 million gamers in the country, generating a yearly revenue of over €6.2 billion

1.2 Project Objectives

Project objectives include assembling and acknowledging the views on online games and gaming habits held by students at a Vietnamese-German university. This survey is aimed at collecting data to determine the online gaming status of Vietnamese-German University students based on a number of outstanding characteristics such as gender, year of study, playing time, the reason for playing games, as well as equipment for this type of entertainment. This also helps to understand more about the gaming habits of Vietnamese-German students, as well as to support and help Vietnamese-German university students to have scientific time management methods and choose the appropriate gaming types.

1.3 Applications

In this project, we discussed how to use a method called hierarchical clustering analysis, which is an algorithm that groups similar objects into groups called clusters. This method is based on a dendrogram to work out the best way to allocate objects to clusters. Clustering is an important tool when it comes to understanding big data and its potential. It can be helpful in many ways, as it can help to group similar objects together, share similar characteristics, and keep a focus on a wide variety of data and this method is important to be aware of the different types of clustering, their purposes, and how to use them.

Chapter 2: Background and State of the art

2.1 What is cluster analysis?

One of the most used methods for exploratory data analysis is clustering. People attempt to gain a first sense of their data across all disciplines, from the social sciences to biology to computer science, by recognizing significant groups within the data points. For instance, behavioral psychologists can segment groups of people based on defined categories (such as eye contact, talking gestures or word used in conversation), or by looking at movie viewing history on a certain platform, developers can categorize users into the respective personality groups [3].

In general, cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). In other words, clustering is used to determine whether the data is divided into a number of distinct subgroups, each group representing objects with distinctly different features, and build descriptive statistics.

There are several famous algorithms applied the idea of clustering such as k-means, k-mean++, or k-medoids that are quite popular and widely used. With many algorithms approaching this clustering method, those algorithms are classified into 2 main directions, which are hierarchical clustering and non-hierarchical clustering. However, in the scope of this report, the former would be emphasized and explained, and details of the latter could be found at [4, 5, 6].

2.2 Hierarchical clustering

Hierarchical clustering constructs hierarchical representations, in which clusters at each level of the hierarchy are formed by merging clusters at the level below. They begin with a single-point cluster for each data point in the simple clustering. There are two main approaches for implementing hierarchical clustering:

Agglomerative: Beginning with each point as its own cluster, combine the nearest two clusters at each stage. The method continues until the largest cluster, which contains all smaller clusters and residual data points, has been generated.

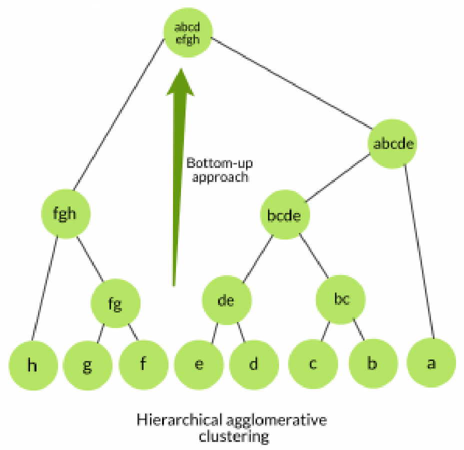


Figure 2.1: Agglomerative clustering.[1]

Divise: Starting with a single, all-inclusive cluster, split one cluster at a time until only singleton clusters of distinct points are left. A strategy for choosing which cluster would divide and how to split is required in this situation.

[1] https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/

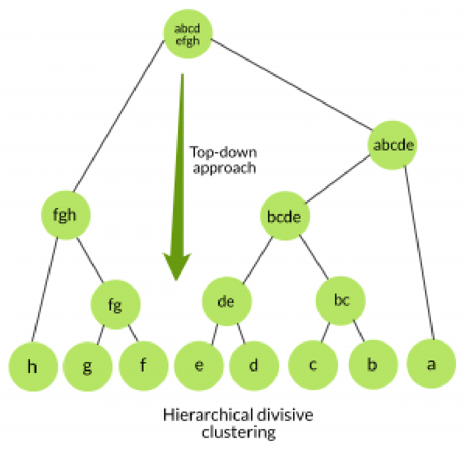


Figure 2.2: Devise clustering.[2]

This research would choose agglomerative clustering because its implementation would be more convenient as well as the literature about it is also richer and more diverse.

2.2.1 Agglomerative clustering

The techniques used in hierarchical clustering are agglomerative in the sense that they begin with entirely fragmented data and continue to construct larger and larger clusters as they go. In detail, the algorithm begins with the simple single-point grouping of data points, which is the most basic form of clustering. The ”closest” groups from the prior clustering are then repeatedly merged using these techniques. With each subsequent round, fewer clusters are created. If these techniques were to be used indefinitely, they would eventually produce trivial clustering, in which every domain point would be part of a single huge cluster. The results of such an algorithm can be represented by a clustering dendrogram, which is a tree of domain subsets with the singleton sets in its leaves and the entire domain at its root, without the need for a stopping rule. The number of clusters to use can be reasonably selected based on the dendrogram which is generated after running the algorithm.

[2] https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/

2.2.2 Distances[3]

There are several measurements for calculating distances between variables. The methods listed below are well-known and widely used methods, there are many more methods mentioned in [6]

Euclidean: This is a well-known formula for measuring the distance between two points, this is the sum of the squared deviations between the measurements for each variable.

$$dist(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

Minkowski: Here, the square is substituted by increasing the difference by a power of m, and the $m^t h$ root is used in place of the square root.

$$dist(x,y) = \sum_{i=1}^{p} |x_i - y_i|^{m^{\frac{1}{m}}}$$

2.2.3 Dissimilarities[4]

Dissimilarities between variables can be calculated using a variety of metrics. The following techniques are often employed techniques.

Single-linkage: Suppose there are two clusters A and B, this method chooses the closest (least dissimilar) pair's intergroup dissimilarity as the standard:

$$d(A,B) = \min\{d(x,y) : x \in A, y \in B\}$$

Complete-linkage: Suppose there are two clusters A and B, the distance between the two clusters is defined as the maximum distance between their element:

$$d(A,B) = \max\{d(x,y) : x \in A, y \in B\}$$

[3] The notation of $dist(x,y)$ stands for the distance from x to y.

Average-linkage: Suppose there are two clusters A and B, the average distance between a point in one cluster and a point in the other is used to quantify the distance between two clusters:

$$d(A,B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x,y)$$

Ward-linkage: The concept is quite similar to the analysis of variance (ANOVA). The rise in the ”error sum of squares” (ESS) after merging two clusters into a single cluster is how the linkage function, which specifies the distance between two clusters, is calculated. At each step of merging, the pairs would be collected so that the ESS would reach the as small as possible value. Let X_{ijk} denote the value for variable k in observation j belonging to cluster i:

$$ESS = \sum_i \sum_j \sum_k |X_{ijk} - \overline{x}_{ik}|^2$$

[4] The notation of $d(A,B)$ stands for the dissimilarity between cluster A and cluster B.

Chapter 3: Methodology

3.1 Data collection

Data collection was done using google form with 7 multiple choice questions. The purpose of this survey was to get VGU student's opinions on their game. The survey included a total of 117 responses and was summarized into 7 pie charts. All details of the survey form can be found at the following link: https://forms.gle/5wC3wqmnZoCzMDqg9

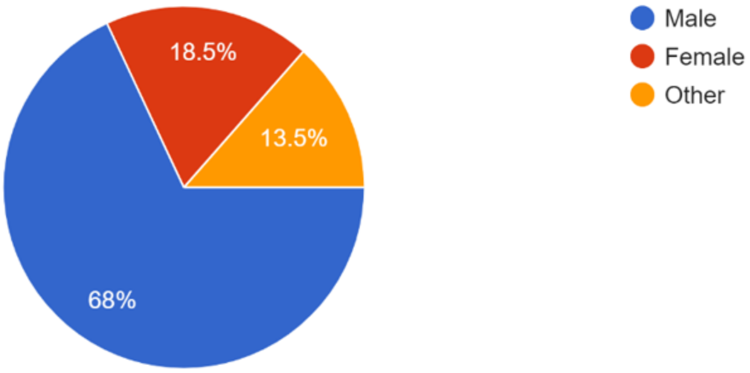


Figure 3.1: Gender ratio of students playing games.

The first chart shows the sex ratios of the participants. Since women are more interested in activities such as shopping or cooking, the percentage of men playing games is absolutely overwhelming.

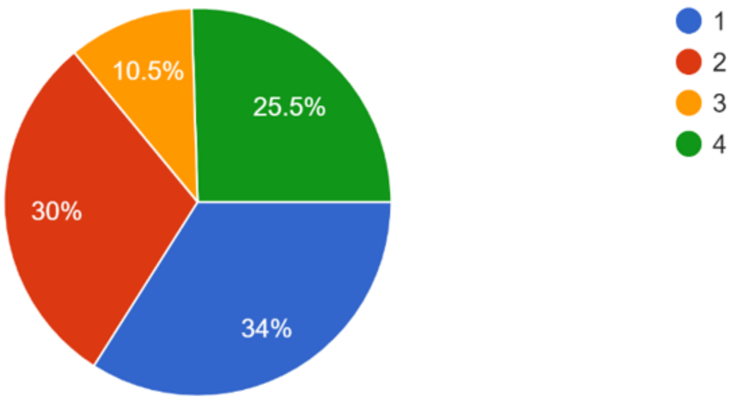


Figure 3.2: Distribution rate of students playing games by year of university.

The second chart shows the percentage of students playing games by years of study in college. The number of students playing games in each year is proportional to the amount of study in the respective years.

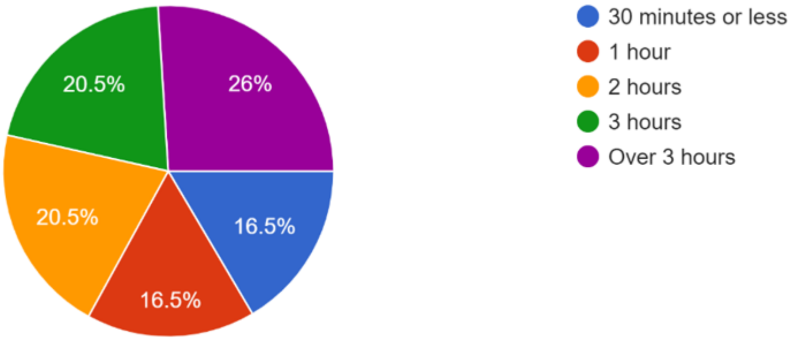


Figure 3.3: Amount of time that students spend playing games in a day.

The third chart shows the student's gaming time in a day. Most students spend 2 to 3 hours a day playing games, the percentage of students playing games for more than 3 hours is also quite high.

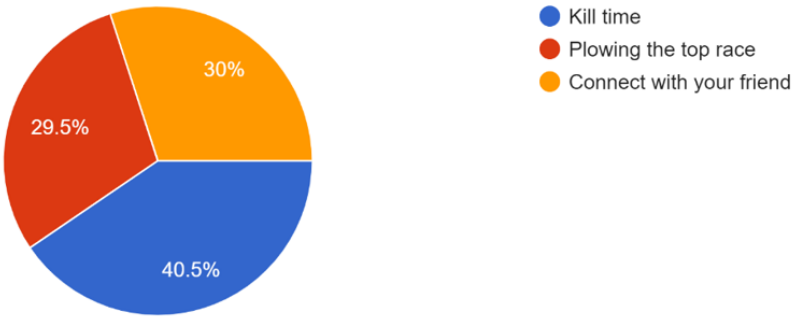


Figure 3.4: Main reason that students play games.

The fourth chart shows why students play games. The three main reasons listed are killing time, competing for the best in the game, and having fun with friends in almost equal proportions.

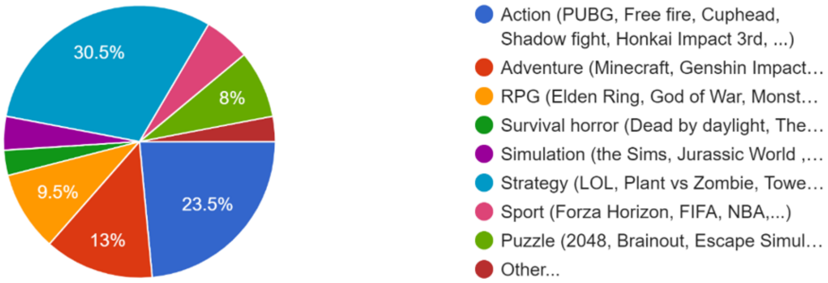


Figure 3.5: Types of games that students usually play.

The fifth chart shows the type of game the students played. Similar to the gaming gender ratio, the action, adventure, and strategy game genres are favored by the majority of players.

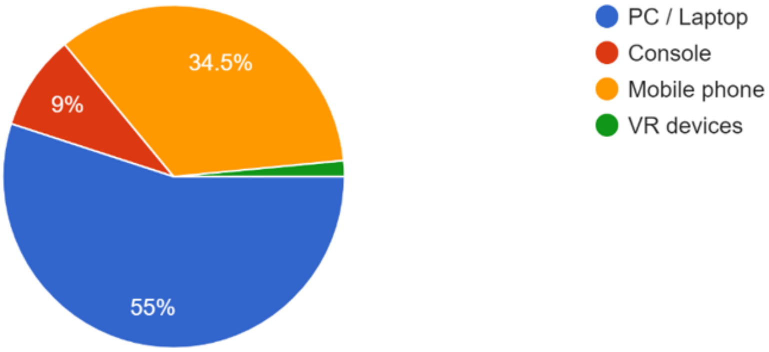


Figure 3.6: Devices that students use to play games.

The sixth chart shows the devices students use to play games. Along with the needs for learning, personal computers, laptops and phones are also convenient for students to play games instead of investing in consoles or virtual reality devices.

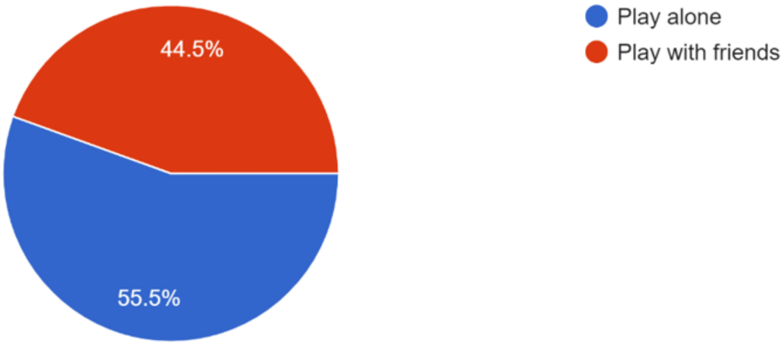


Figure 3.7: Students playing games alone and playing games with friends.

The last chart shows the odds of playing alone and playing with friends. The percentage of VGU students playing games alone is slightly higher than those playing with friends

3.2 System architecture

The flow of how our research conduct is describes as below:

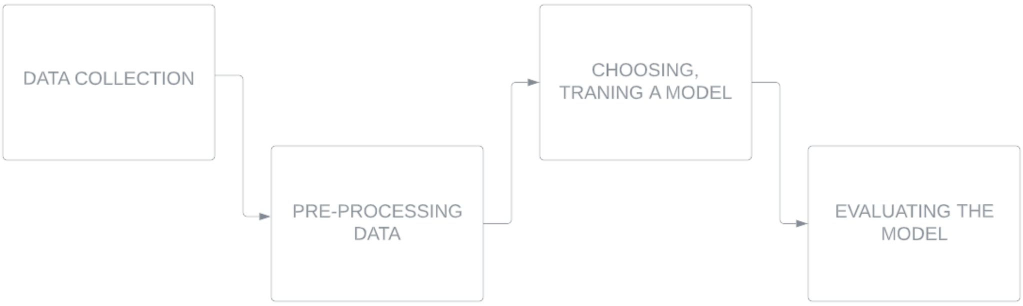
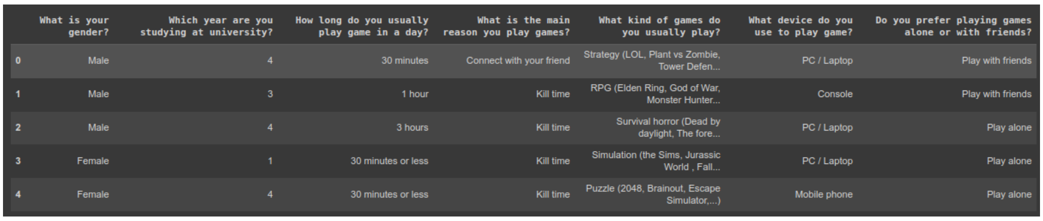


Figure 3.8: The flow of the application.

3.2.1 Pre-processing data

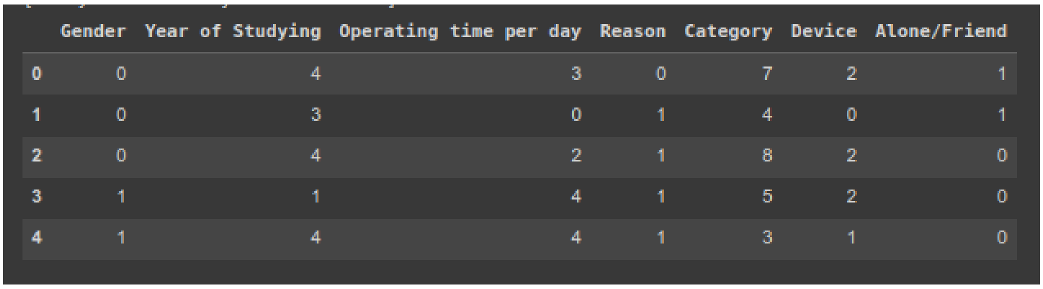
The method of obtaining information and the parameters of the data collected are described in detail in section 3.1. Multiple choice questions make it easier to collect data when the answer is clear, accurate, and more effective than if the response is in the form of sentences, the conclusion of a specific intelligence in those descriptive sentences describes description will not be as transparent. Responses, after being gathered, are converted into a comma-separated values (CSV) form. As can be seen from the figure below, responses are still in the form of categorical data:



	What is your gender?	Which year are you studying at university?	How long do you usually play game in a day?	What is the main reason you play games?	What kind of games do you usually play?	What device do you use to play game?	Do you prefer playing games alone or with friends?
0	Male	4	30 minutes	Connect with your friend	Strategy (LOL, Plant vs Zombie, Tower Defen...	PC / Laptop	Play with friends
1	Male	3	1 hour	Kill time	RPG (Elden Ring, God of War, Monster Hunter...	Console	Play with friends
2	Male	4	3 hours	Kill time	Survival horror (Dead by daylight, The fore...	PC / Laptop	Play alone
3	Female	1	30 minutes or less	Kill time	Simulation (the Sims, Jurassic World , Fall...	PC / Laptop	Play alone
4	Female	4	30 minutes or less	Kill time	Puzzle (2048, Brainout, Escape Simulator,...)	Mobile phone	Play alone

Figure 3.9: Responses in a form of categorical data.

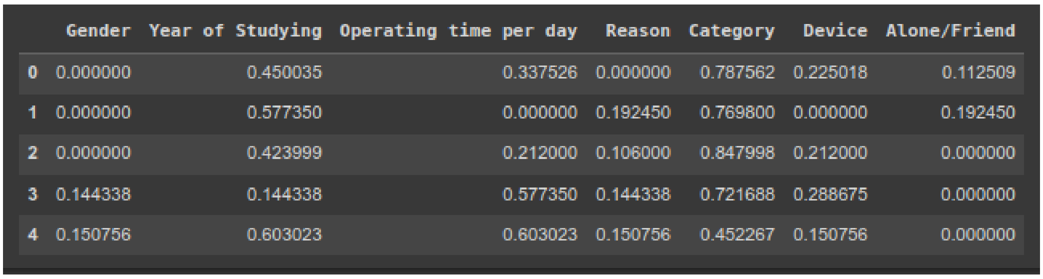
Keeping such values can make the data transparent to the reader, however, algorithms are almost impossible to deal with if the values are in this form. There are several methods for handling this problem, such as creating dummy variables (refer the actual values of the data to another value domain [7]), stratification (A dataset is stratified by dividing it into smaller groups based on the results of a category variable) and other techniques of encoding [8]. The research uses a method of creating dummy variables for converting categorical data into a numerical form.



	Gender	Year of Studying	Operating time per day	Reason	Category	Device	Alone/Friend
0	0	4	3	0	7	2	1
1	0	3	0	1	4	0	1
2	0	4	2	1	8	2	0
3	1	1	4	1	5	2	0
4	1	4	4	1	3	1	0

Figure 3.10: Create dummy variables.

Notably, the values within each feature lie on very different ranges, for instance, values of 'Gender' in Figure 3.10 are mostly seen as 0 and 1, the phenomenon is identical in the case of 'Device' category when the data to be seen is 0, 1 and 2. Indeed, that value domain is really different, more specifically, for each question given, the number of answers for each of those questions is initialized with a different number, after creating dummy variables, the value domain of each question (each feature) will differ by different degrees (number of answers per question). Normalization would be an appropriate approach to this phenomenon. This action could prevent outliers as well as convert the range of the attributes to the most common form in the range [0, 1]. Also, we use Euclidean distance for measuring the distance between data points and clusters. All the measurements for calculating dissimilarities discussed in section 2.2.3 are used.



	Gender	Year of Studying	Operating time per day	Reason	Category	Device	Alone/Friend
0	0.000000	0.450035	0.337526	0.000000	0.787562	0.225018	0.112509
1	0.000000	0.577350	0.000000	0.192450	0.769800	0.000000	0.192450
2	0.000000	0.423999	0.212000	0.106000	0.847998	0.212000	0.000000
3	0.144338	0.144338	0.577350	0.144338	0.721688	0.288675	0.000000
4	0.150756	0.603023	0.603023	0.150756	0.452267	0.150756	0.000000

Figure 3.11: Doing normalization.

3.2.2 Choosing and training a model

The research investigates the behavior of VGU students on gaming habits using hierarchical clustering. As mentioned in section 2.2, this research is conducted using agglomerative clustering and several methods for calculating dissimilarities are introduced in section 2.2.3.

3.2.3 Evaluating the model

After doing the training session, we realize that applying other techniques produce better outcome. In detail, as a prerequisite step before performing the training process, applying principle component analysis (PCA) removes unnecessary features, making clustering more intuitive.

Principle component analysis (PCA) is known as a technique of dimensionality reduction, using this method, a large set of variables is able to transform into a smaller one that still contains most of the information in the large set [?]. We want to choose suitable features that, when performing clustering, will give clearly separated values. Using PCA for evaluation, we use standardization instead of normalization at the pre-processing step. This stage standardizes the range of the continuous starting variables with the intention of ensuring that each one contributes equally to the analysis.

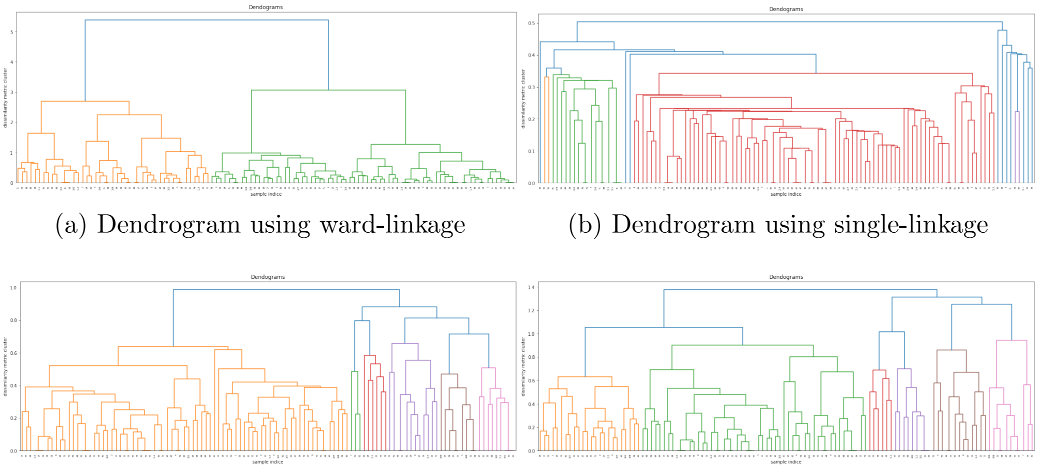
Chapter 4: Results

4.1 Results after training

In case of having more data with more extreme outliers, we also propose another preprocessing data method normalizing in order to enhance the performance and reliability of algorithm. Normalization is a scaling technique that changes the values of numeric columns in the dataset to use a common scale. In our case study, we used a method called Normalization, which helps the dataset to shift and rescale the values of their attributes, so they end up ranging between 0 and 1. It is useful when feature distribution is unknown. And these is our results using Normalized data.

4.1.1 Dendrograms

After performing data preprocessing, to determine the number of clusters needed to perform clustering, we use a dendrogram to describe the relationship between data points as well as clusters. All the method of measuring linkage are described in section 2.2.3



(a) Dendrogram using ward-linkage

(b) Dendrogram using single-linkage

(c) Dendrogram using average-linkage

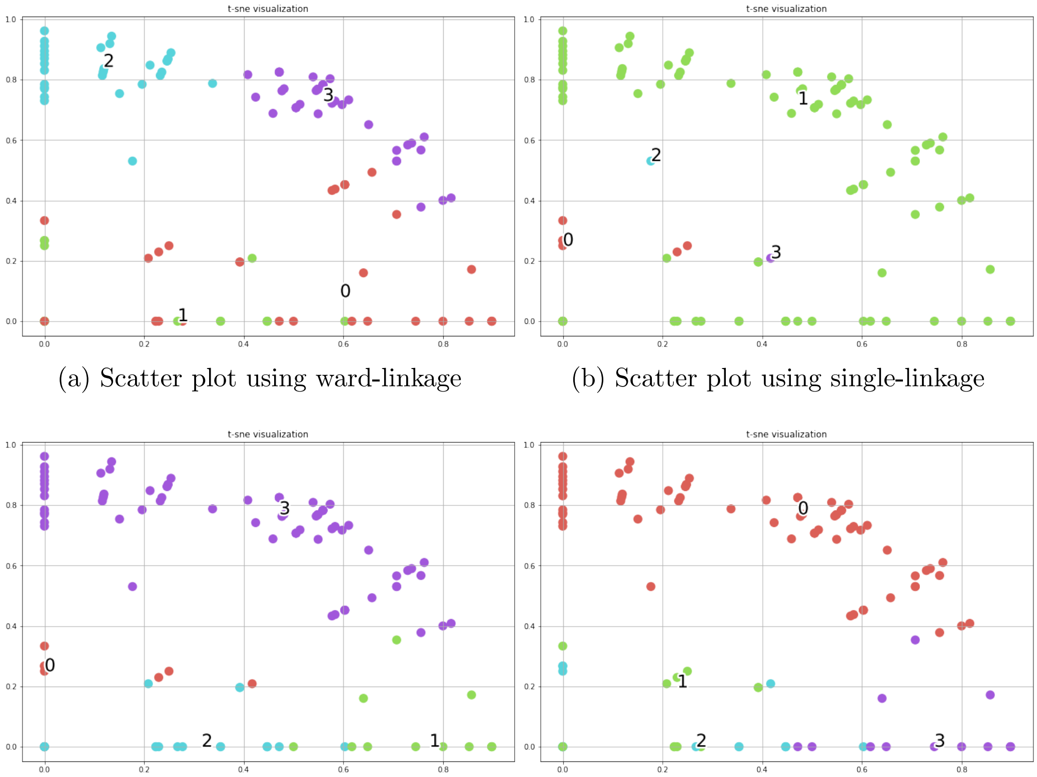
(d) Dendrogram using complete-linkage

Figure 4.1: Dendrograms after pre-processing

A dendrogram is a diagram that depicts the relationship between things in terms of hierarchy. A dendrogram is mostly used to determine how to assign items to clusters. Single-linkage measure the dissimilarities based choosing the closest (least dissimilar) pair's intergroup dissimilarity, therefor, it will often mix observations connected by a string of closely spaced intermediate observations at relatively low thresholds (Figure. 4.1b). Meanwhile, complete-linkage, in details, a cluster's allocated observations may be substantially closer to other cluster members than they are to some of its own (Figure. 4.1d). An average-linkage compromises single-linkage and average-linkage(Figure. 4.1c)[5]. Instead of employing association or distance metrics, ward-linkage approaches cluster analysis as an analysis of variance problem, the results are clearer than the other methods (Figure. 4.1a After visualize the relationship between clusters, we reached the conclusion that the sufficient number of clusters is 4.

4.1.2 Scatter plot after clustering

After training, we get the result and perform it on the graph. The graph below is redrawn with the number of clusters for all the methods mentioned in section 2.2.3 is 4.



(a) Scatter plot using ward-linkage

(b) Scatter plot using single-linkage

(c) Scatter plot using average-linkage

(d) Scatter plot using complete-linkage

Figure 4.2: Dendrograms after pre-processing

As can be seen from the figure, the cluster part is close to achieving the best state when it is possible to classify most of the data points, however, there are still some points that are doped among the clusters. We can see that Figure. 4.2b with cluster 1 occupies most of the data points. The remaining figures have similarity when the data points are clearly distinguished between clusters except in clusters 0 and 1 (Figure. 4.2a), clusters 1 and 2 (Figure. 4.2c) and clusters 1, 2 and 3 in Figure. 4.2d.

4.2 Results after evaluation

The dendrogram of hierarchical agglomerative clustering is shown in Fig There are so many clusters formed since the hierarchical clustering doesn't need the number of clusters to be prefixed.

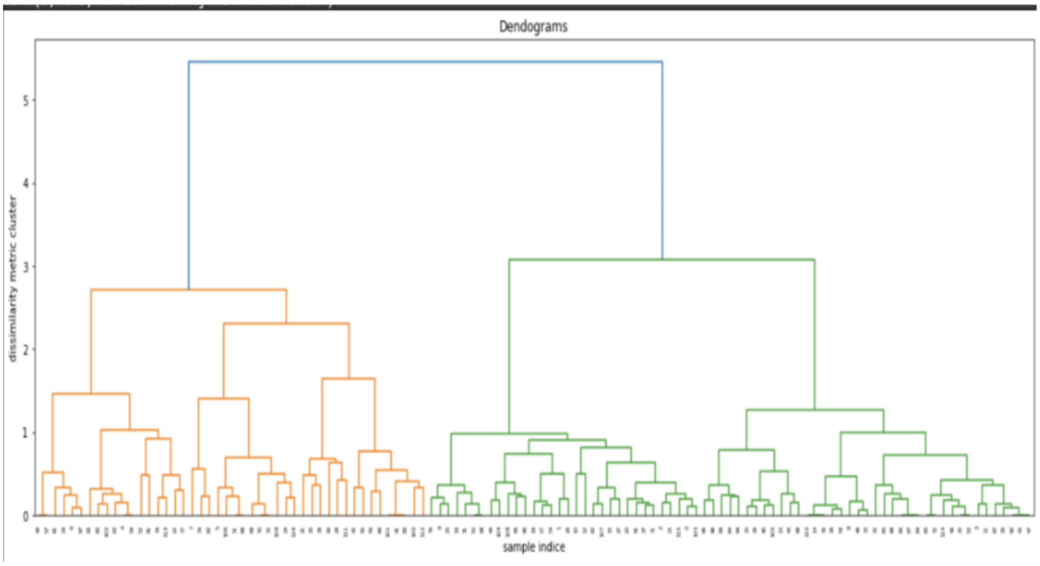
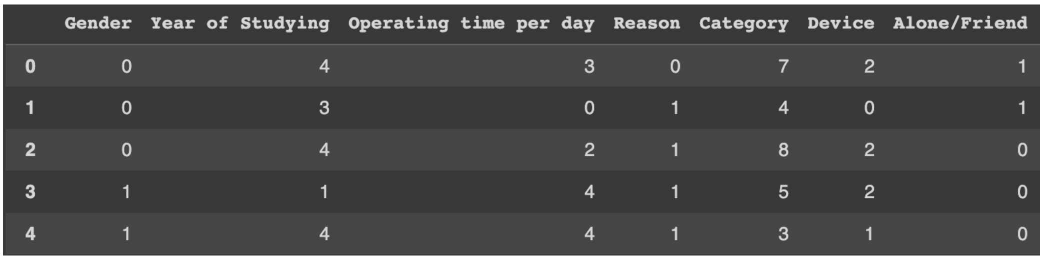


Figure 4.3: Dendrogram of gaming habit in VGU

4.2.1 Dataset Description

The proposed work is analyzed with a variant sector namely Gaming habits. This Dataset was taken by our team, using Google Surveys in 4 days. It includes 117 instances and seven attributes. We selected the attribute that may affect the habit, due to the lack of time we can't have a proper consideration of questions or measure the correctness of our data, which may lead to potential risks in our analysis. For example, we couldn't conduct the weight of each parameter, which can mislead us to extract wrong insight.



	Gender	Year of Studying	Operating time per day	Reason	Category	Device	Alone/Friend
0	0	4	3	0	7	2	1
1	0	3	0	1	4	0	1
2	0	4	2	1	8	2	0
3	1	1	4	1	5	2	0
4	1	4	4	1	3	1	0

Figure 4.4: Dataset

4.2.2 Results

In the end, we conducted three plots each of them using a different method of linkage (Ward, Single, Complete).

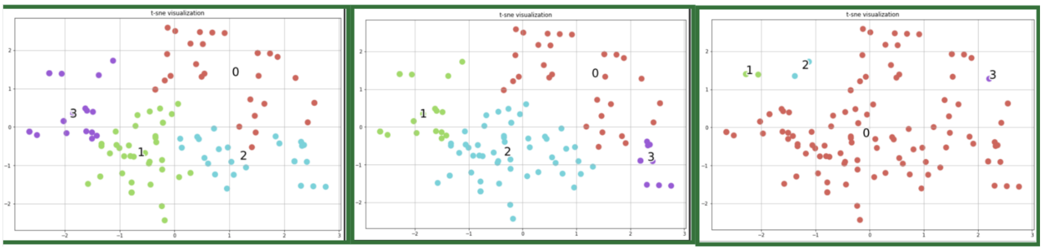


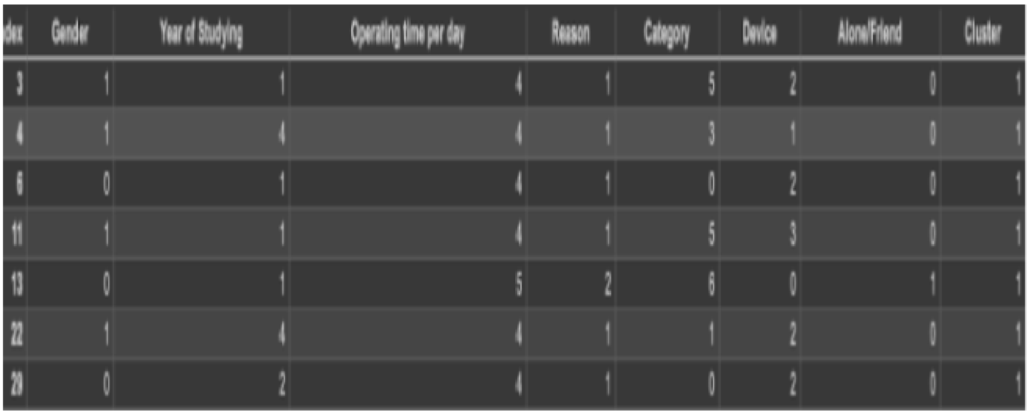
Figure 4.5: Clustering using Ward, Complete, Single

Each of them gave us a different result. The most abnormal plot is using the Single method, which has the most data belonging to cluster 0 and can not lead us to any of the insight. This problem can be occurred by having a tendency to combine, at relatively low thresholds, observations linked by a series of close intermediate observations. This phenomenon, referred to as chaining, is often considered a defect of a method. Single is fast and can perform well on non-globular data, but it performs poorly in the presence of noise. Ward is the most effective method for noisy data. As you can see it can be plotted most intuitively. Complete also gave us a fairly intuitive plot but it has a few points different from than Ward method. To draw out the conclusion, we decided to use Ward as a result of our analysis.

Chapter 5: Conclusion and future works

5.1 Conclusion

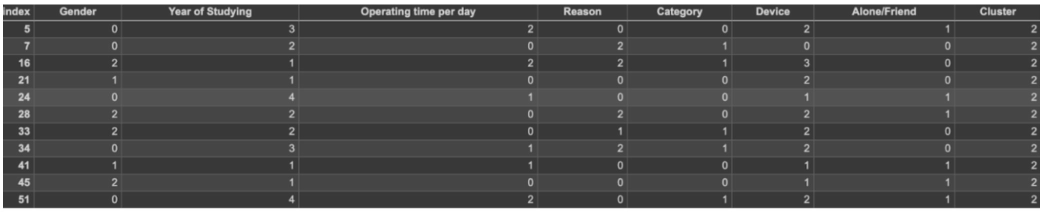
The data collected have been clustered into different groups. A conclusion has been conducted.



dex	Gender	Year of Studying	Operating time per day	Reason	Category	Device	Alone/Friend	Cluster
3	1	1	4	1	5	2	0	1
4	1	4	4	1	3	1	0	1
6	0	1	4	1	0	2	0	1
11	1	1	4	1	5	3	0	1
13	0	1	5	2	6	0	1	1
22	1	4	4	1	1	2	0	1
29	0	2	4	1	0	2	0	1

Figure 5.1: Grouped data

With cluster 1, students have play time above 4 hours often playing with their friends and the reason to play is to kill time, they are a group having the most time to play with a variety of playing devices and game categories not focusing on any specific type of game. Mostly depend on their friends. With the population mainly freshmen and seniors.



ndex	Gender	Year of Studying	Operating time per day	Reason	Category	Device	Alone/Friend	Cluster
5	0	3	2	0	0	2	1	2
7	0	2	0	2	1	0	0	2
16	2	1	2	2	1	3	0	2
21	1	1	0	0	0	2	0	2
24	0	4	1	0	0	1	1	2
28	2	2	0	2	0	2	1	2
33	2	2	0	1	1	2	0	2
34	0	3	1	2	1	2	0	2
41	1	1	1	0	0	1	1	2
45	2	1	0	0	0	1	1	2
51	0	4	2	0	1	2	1	2

Figure 5.2: Grouped data

For cluster 2, we're having the group with less playing time, mostly from 0 - 2 hours per day. And most of the time, they're playing games with the category Action and Adventure. The percentage of VGUers in this group having pretty equal between playing with a friend or not so can give us insight into this group unlike cluster 0, which heavily depends on their friends.

So on the other group, a few features were extracted from the hierarchical clustering algorithm with 2 other clusters. Male students are having more time spending on games than females and they are mainly playing action games, with the highest amount of time. While female students have less playing time but would rather join with friends. Freshmen and seniors are taking 1st place in playing times. In conclusion, after running through the analysis, we can summarize that friends are the heaviest parameter that can affect the gaming habit of VGU students, and in the population of our survey action games are also attached with the highest amount of playing time group. Finally, it might be too early to draw a conclusion on occupation as some other features have a strong impact on gaming habit preference. Additionally, for time limitations, we skip the robustness check step. It could hurt the validity of our clustering results.

5.2 Future work

Selecting an optimal number of clusters is key to applying a clustering algorithm to the dataset, such as k-means clustering, which requires the user to specify the number of clusters k to be generated. This is a somewhat arbitrary procedure, one of the weakest aspects of performing cluster analysis. The Elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. It works by finding WCSS (Within-Cluster Sum of Square) i.e. the sum of the square distance between points in a cluster and the cluster centroid.

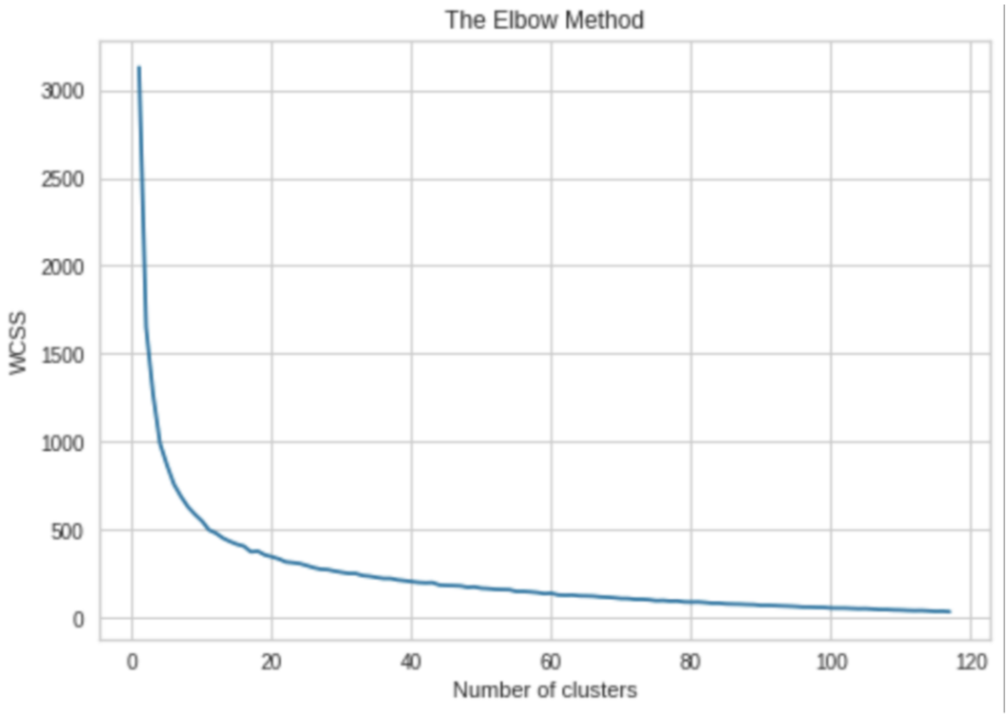


Figure 5.3: Elbow method 1-117 clusters

This is the result after we ran the elbow method which lead us to the conclusion of how many clusters we should have. The reason why we got k = 118, is because we got a total of 117 responses, and any of that can be affected the number of clusters As the figure, we can see that the slope of the line is drastically changed at 0-10 clusters. In order to go deeper and more specific about how many clusters,

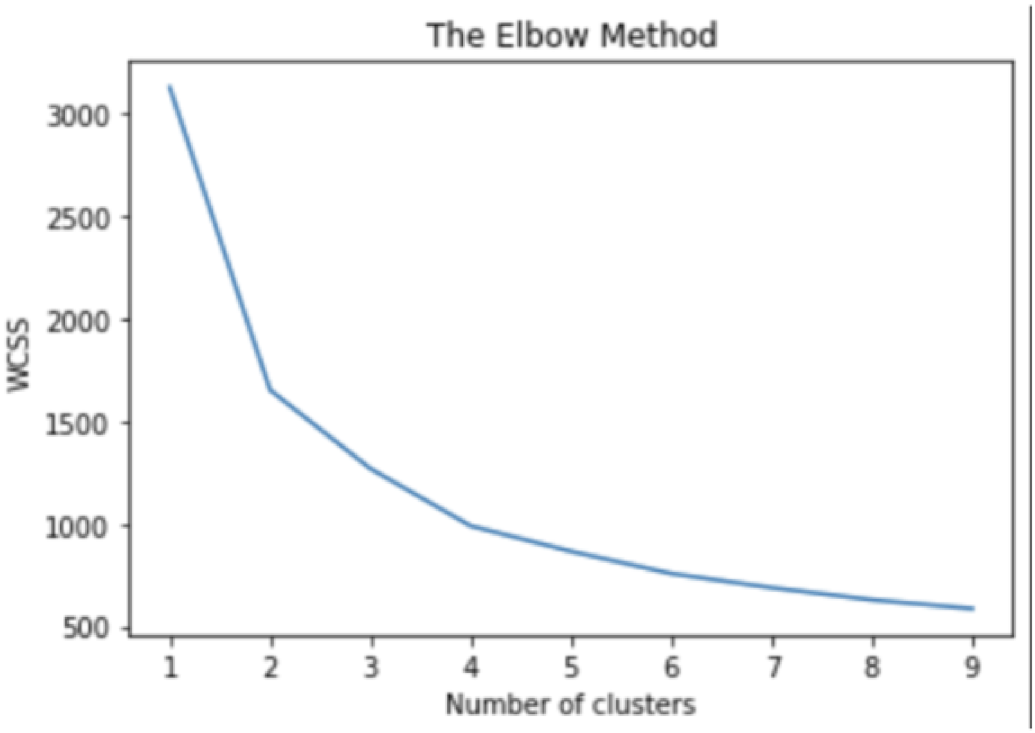
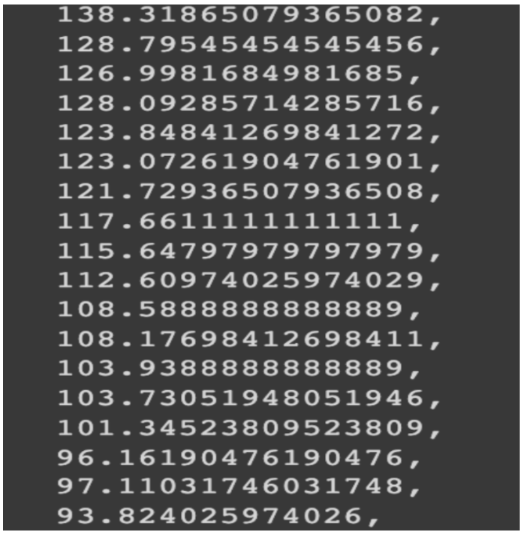


Figure 5.4: Elbow method 1-9 clusters

we ran another plot which include 1-9 clusters. As you can see that the plot shows which of the cure are elbow, there are two slopes we need to pay attention to, which are 2 and 4 clusters, but 2 clusters won't give us insight of the data and skip many important features of the data we get. After coming up with that conclusion, we ran all the hierarchical clustering algorithm with a number of clusters is 4. To prove that our method is not in the plot but also in numbers this is the WCSS (Within-Cluster Sum of Square) of our data, the sum of the square distance between points in a cluster and the cluster centroid.

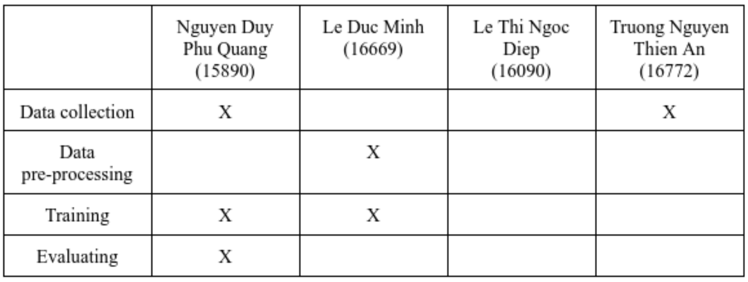


```
138.31865079365082,
128.79545454545456,
126.9981684981685,
128.09285714285716,
123.84841269841272,
123.07261904761901,
121.72936507936508,
117.6611111111111,
115.64797979797979,
112.60974025974029,
108.5888888888889,
108.17698412698411,
103.9388888888889,
103.73051948051946,
101.34523809523809,
96.16190476190476,
97.11031746031748,
93.824025974026,
```

Figure 5.5: WCSS index

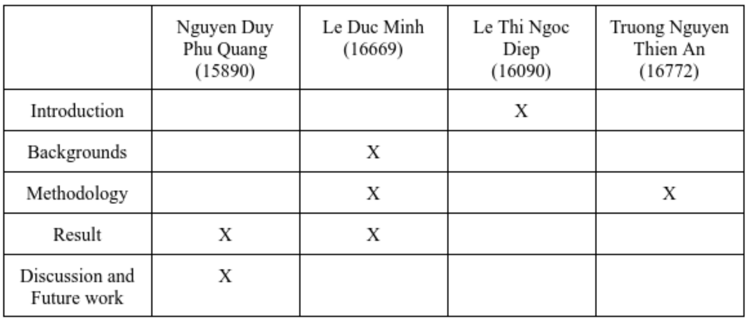
Appendix A: Contribution

This is a record of the amount of work each individual does during the project implementation.



	Nguyen Duy Phu Quang (15890)	Le Duc Minh (16669)	Le Thi Ngoc Diep (16090)	Truong Nguyen Thien An (16772)
Data collection	X			X
Data pre-processing		X		
Training	X	X		
Evaluating	X			

Figure A.1: Contributing in code implementation



	Nguyen Duy Phu Quang (15890)	Le Duc Minh (16669)	Le Thi Ngoc Diep (16090)	Truong Nguyen Thien An (16772)
Introduction			X	
Backgrounds		X		
Methodology		X		X
Result	X	X		
Discussion and Future work	X			

Figure A.2: Contributing in writing document

Appendix B: Code and Dataset

This is our code that was made during the course of the research.

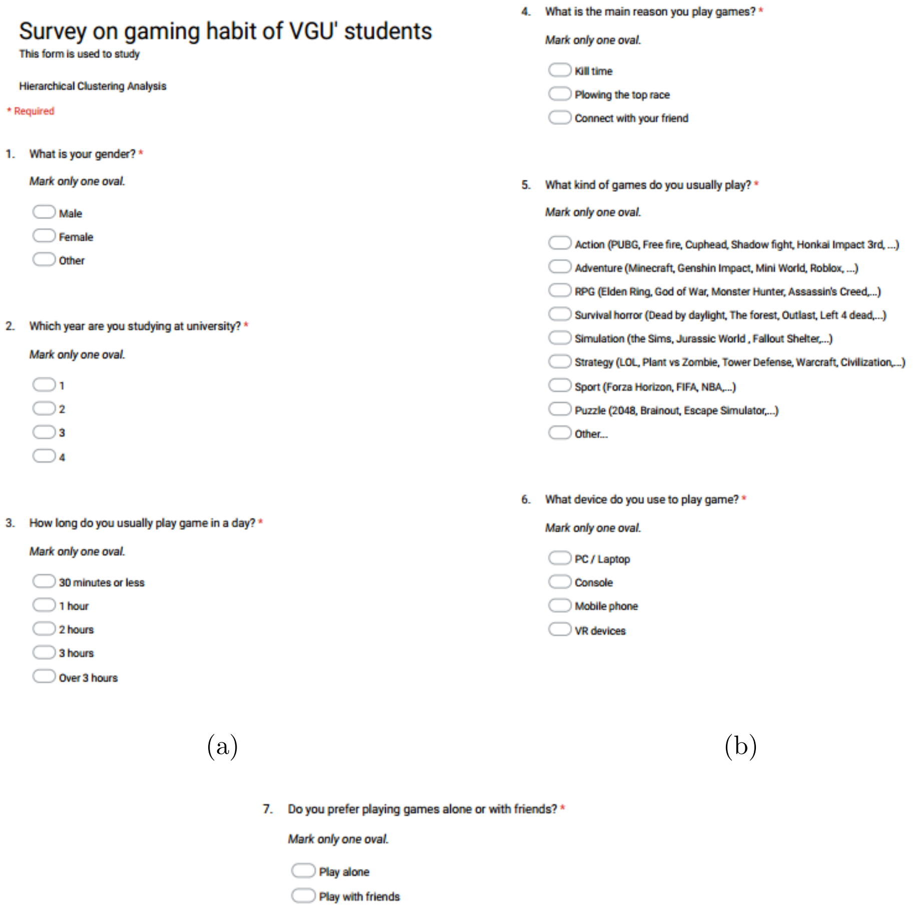
https://colab.research.google.com/drive/13TYx-7eKg4XoWET0hLgfPK-1ajBcaIyk?usp=sharing

These are responses that were collected during the survey.

https://github.com/MinLee0210/-VGU-HDDA-Survey-responses.git

Appendix C: Survey

This is the content and layout of our survey.



Survey on gaming habit of VGU' students

This form is used to study

Hierarchical Clustering Analysis

* Required

1. What is your gender? *

   Mark only one oval.

   ○ Male
   ○ Female
   ○ Other

2. Which year are you studying at university? *

   Mark only one oval.

   ○ 1
   ○ 2
   ○ 3
   ○ 4

3. How long do you usually play game in a day? *

   Mark only one oval.

   ○ 30 minutes or less
   ○ 1 hour
   ○ 2 hours
   ○ 3 hours
   ○ Over 3 hours

(a)

4. What is the main reason you play games? *

   Mark only one oval.

   ○ Kill time
   ○ Plowing the top race
   ○ Connect with your friend

5. What kind of games do you usually play? *

   Mark only one oval.

   ○ Action (PUBG, Free fire, Cuphead, Shadow fight, Honkai Impact 3rd, ...)
   ○ Adventure (Minecraft, Genshin Impact, Mini World, Roblox, ...)
   ○ RPG (Elden Ring, God of War, Monster Hunter, Assassin's Creed,...)
   ○ Survival horror (Dead by daylight, The forest, Outlast, Left 4 dead,...)
   ○ Simulation (the Sims, Jurassic World , Fallout Shelter,...)
   ○ Strategy (LOL, Plant vs Zombie, Tower Defense, Warcraft, Civilization,...)
   ○ Sport (Forza Horizon, FIFA, NBA,...)
   ○ Puzzle (2048, Brainout, Escape Simulator,...)
   ○ Other...

6. What device do you use to play game? *

   Mark only one oval.

   ○ PC / Laptop
   ○ Console
   ○ Mobile phone
   ○ VR devices

(b)

7. Do you prefer playing games alone or with friends? *

   Mark only one oval.

   ○ Play alone
   ○ Play with friends

(c)

Figure C.1: The content and layout of the survey.

References

[1] Mikko Meriläinen. Young people's engagement with digital gaming cultures – validating and developing the digital gaming relationship theory. *Entertainment Computing*, 44:100538, 11 2022.

[2] Adam S. Kahn et al. The trojan player typology: A cross-genre, cross-cultural, behaviorally validated scale of video game play motivations. *Computers in Human Behavior*, 49:354–361, 2015.

[3] Alexandra Roshchina, John Cardiff, and Paolo Rosso. Twin: Personality-based intelligent recommender system. *Journal of Intelligent & Fuzzy Systems*, 28:2059–2071, 06 2015. https://www.researchgate.net/publication/281198129_TWIN_Personality-based_Intelligent_Recommender_System.

[4] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. volume 8, pages 1027–1035, 01 2007.

[5] Trevor Hestie, Robert Tibshirani, and Jerome Friedman. *The Elements Of Statistical Learning: Data Mining, Inference, And Prediction*. Statistic. Springer, 2009.

[6] Azizul Hamid Md Doulah, Md. Siraj Hakim. Performance analysis of hierarchical and non- hierarchical clustering techniques. *STM Journals*, 9:54–71, 10 2020. https://www.researchgate.net/publication/344758615_Performance_Analysis_of_Hierarchical_and_Non-_Hierarchical_Clustering_Techniques.

[7] M Venkataramana, M Subbarayudu, Rajani Meejuru, and K Sreenivasulu. Regression analysis with categorical variables. *International Journal of Statistics and Systems*, 11:135–143, 01 2016. https://www.researchgate.net/publication/348421623_Regression_Analysis_with_Categorical_Variables.

[8] Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *CoRR*, abs/1806.00979, 2018. http://arxiv.org/abs/1806.00979.