GAMES AND STUDENTS

Team Members

HIEW LONG CHEN (타일러)	2023320156
AHMAD FAHREEN QUSYAIRI (아마드)	2023320324
NUR IRDINA BINTI ZAHARI (누르 일디나)	2023320029
NUR ASHA SYAFIQAH BINTI SUHAIMI (아샤)	2023320304

Outline

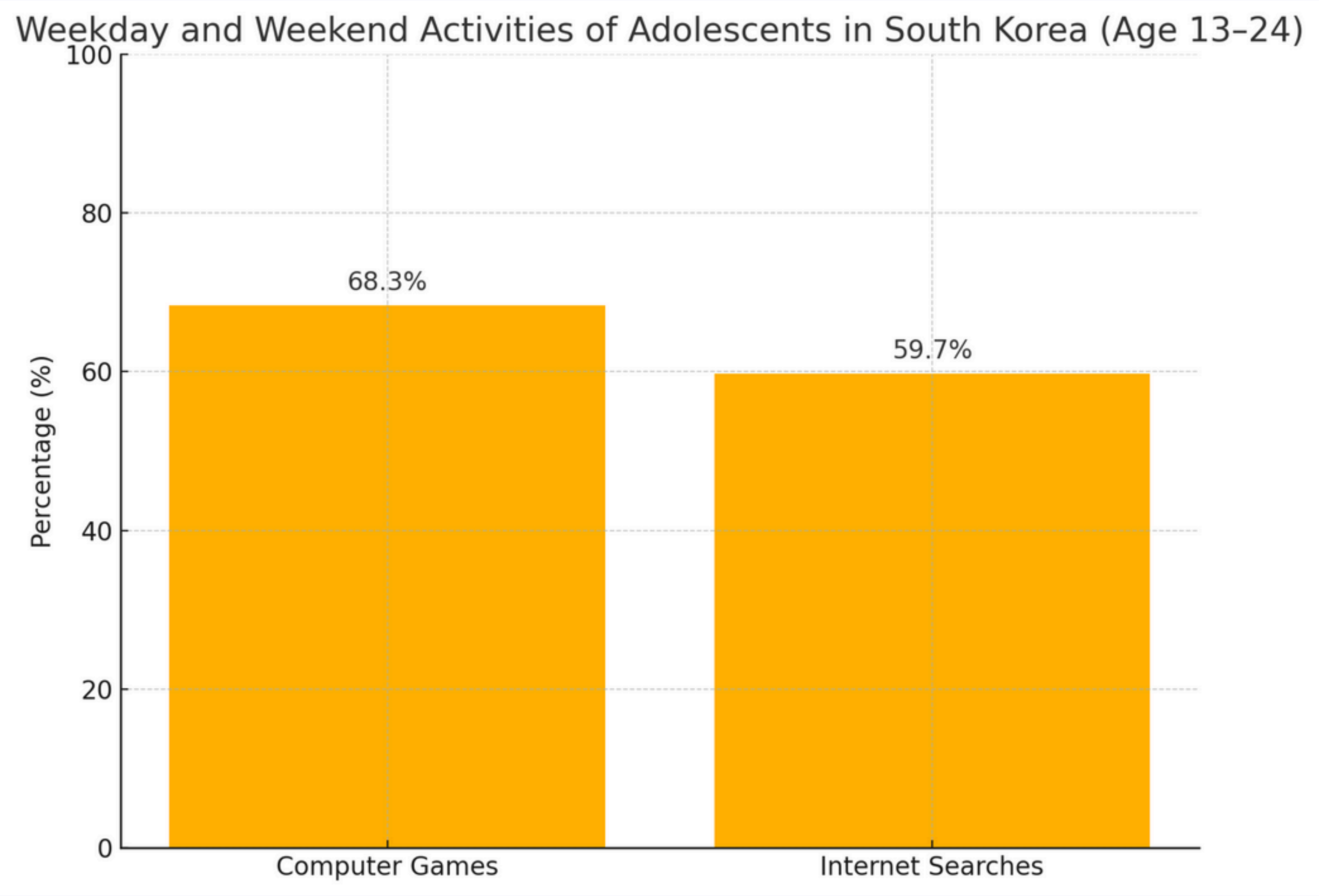
Problem Statement/ Motivation

Goal

Our dataset

Methods

Problem Statement & Motivation



Source: https://doi.org/10.1016/j.childyouth.2018.09.009

Problem Statement

According to a study conducted in 2018, **68.3% of adolescents** (13-24 years old) spend their weekend **playing games.**

Gaming has become a popular trend among students.

Does gaming habits, such as playing time and frequency, have a measurable effect on students' academic performance?

Motivation

Problem Statement & Motivation

Problem Statement

Motivation

As students increasingly engage in gaming, it becomes essential for educators, parents, and students themselves to understand the behavioral patterns that may support or hinder academic success.

Discover the relationships in the dataset that can relate student behavior and their academic performance.

Goal

- **Group students into meaningful clusters** based on their gaming habits and academic performance.
- **Train a classification model to predict a student's performance level** based on gaming and background features.
- **Identify which features** (e.g., playing hours, parent education) are most influential in predicting grades.
- Discover frequent behavior patterns using **pattern mining** (e.g., "High playing hours + low parent education → low grades").
- **Provide interpretable visualizations and insights** to support academic decision-making and self-awareness among students.

Looking into our Dataset

kaggle gameandgrade new.csv

# Sex	# School Code	# Playing Years	# Playing Often	# Playing Hours	# Playing Games	# Parent Revenue	# Father Education	# Mother Education	Grade
1	2	3	4	5	6	7	8	9	10
0 1	1 11	0 4	0 5	0 5	0 2	0 4	0 6	0 6	100 14% 70.00 8% Other (605) 79%
0	1	1	2	1	1	4	4	5	77.50
1	1	1	3	1	1	1	3	3	83.00
0	1	0	0	0	0	1	3	3	80.00
0	1	3	5	1	1	2	2	3	45.00
1	1	1	1	2	1	1	3	4	85.00
0	1	1	5	1	1	1	2	2	80.00
0	1	1	2	2	1	2	3	3	55.00
0	1	1	5	2	1	2	3	3	80.00
1	1	2	1	1	1	3	3	5	60.00
0	1	2	5	2	1	1	2	4	88.00

What data types are we dealing with? How do we process/interpret them?

Our Dataset

770 student records

The dataset records how frequently students engage in gaming, likely measured on a scale (e.g. 1 to 5), providing a quantitative representation of their gaming intensity

Sex	Playing Years	Playing Hours	Parent Revenue	Parent Education	Grade
0	1	1	4	4.5	77.5

Target variable

These variables show how parents' education and family finances could affect their children's gaming habits and grades.

Data Preprocessing (How do we deal with abnormal data?)

Example in dataset:

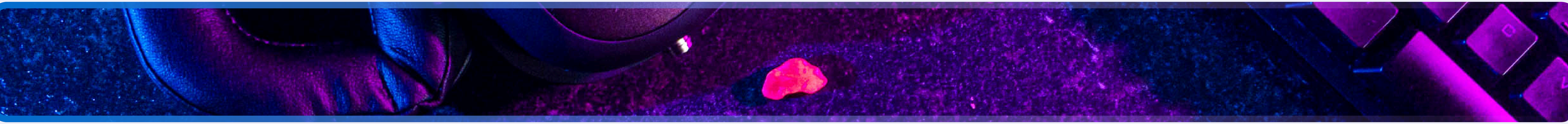
Sex	Playing Years	Playing Hours	Parent Revenue	Father Education	Mother Education	Grade
0	0	0	0	0	0	100

- The entire row is removed to ensure data quality and avoid noise.

Sex	Playing Years	Playing Hours	Parent Revenue	Father Education	Mother Education	Grade
0	0	0	2	4	4	92..00

- Field value is manually corrected to ensure consistent formatting.

Methods



- Clustering
- Classification
- Pattern Mining

Clustering

Objective : To identify the clusters of students with varying gaming habits (e.g. playing hours) and traits (grades).

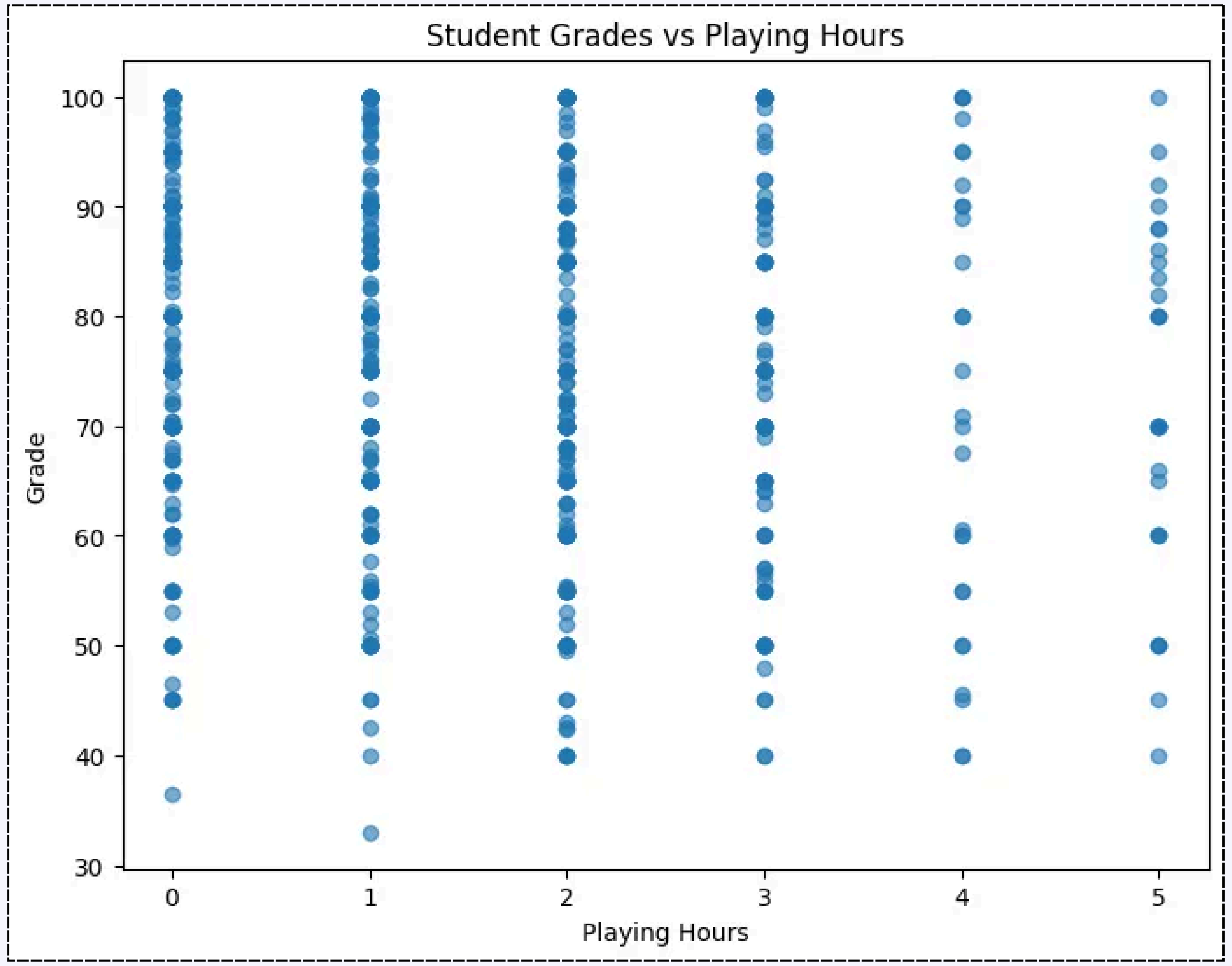
OUR IDEA !

use **K-Means++** as our clustering method approach with :

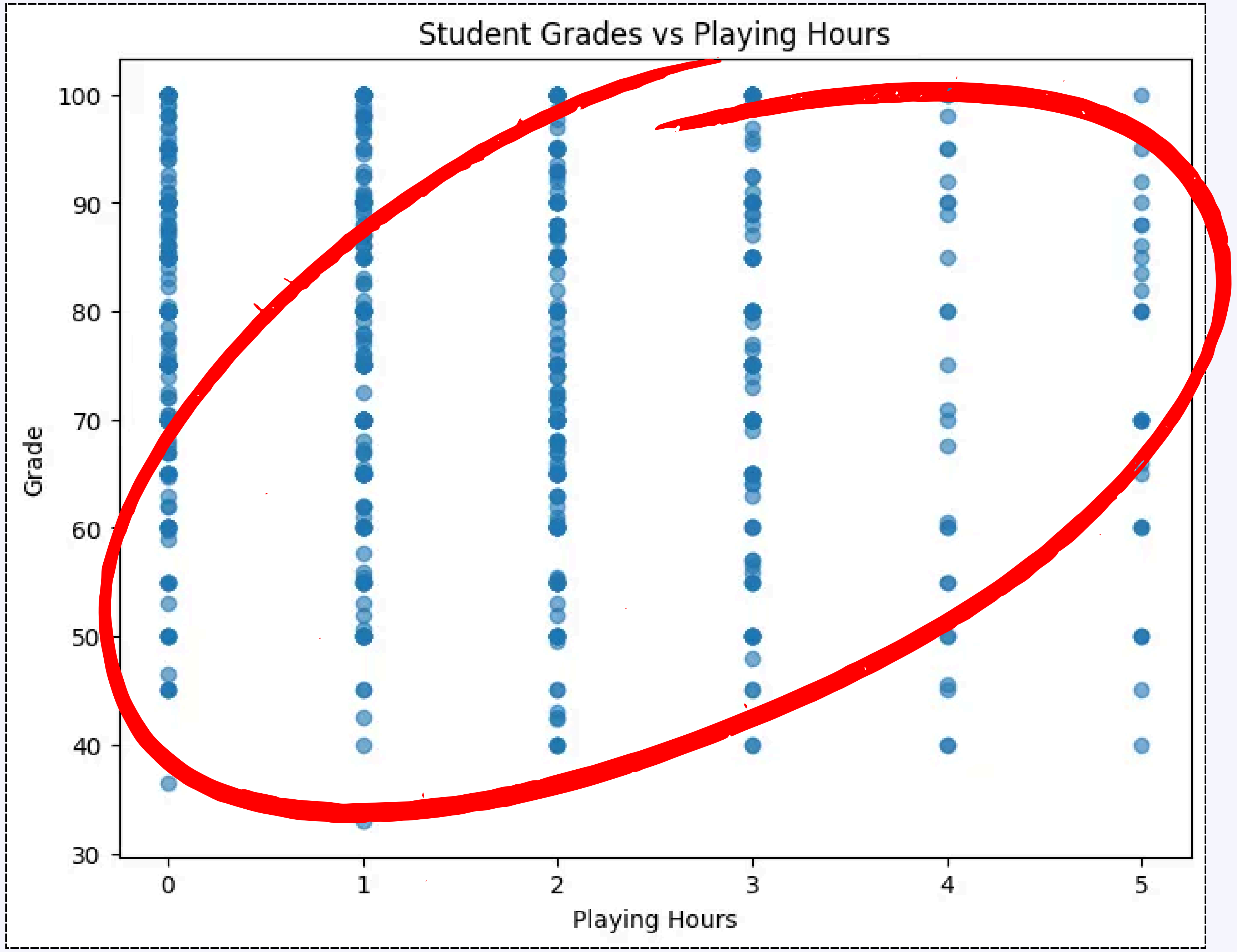
- x-axis : Playing hours |
- Parent Revenue |
  Parent education
- y-axis : Grades

- Graph 1 : Playing hours & Grades
- Graph 2 : Parent Revenue & Grades
- Graph 3 : Parent Education & Grades (in dataset, we have F&M education, so we just take average F+M education)

GRAPH THAT WE GOT AFTER PLOTTING



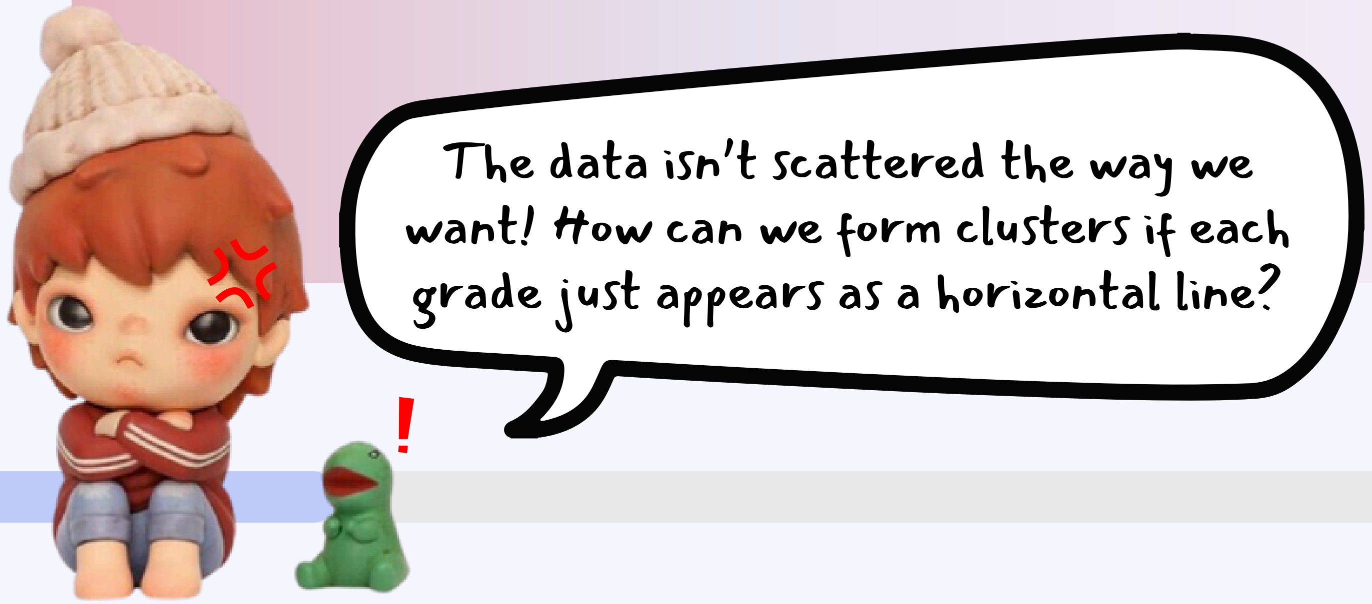
Clustering



Problem

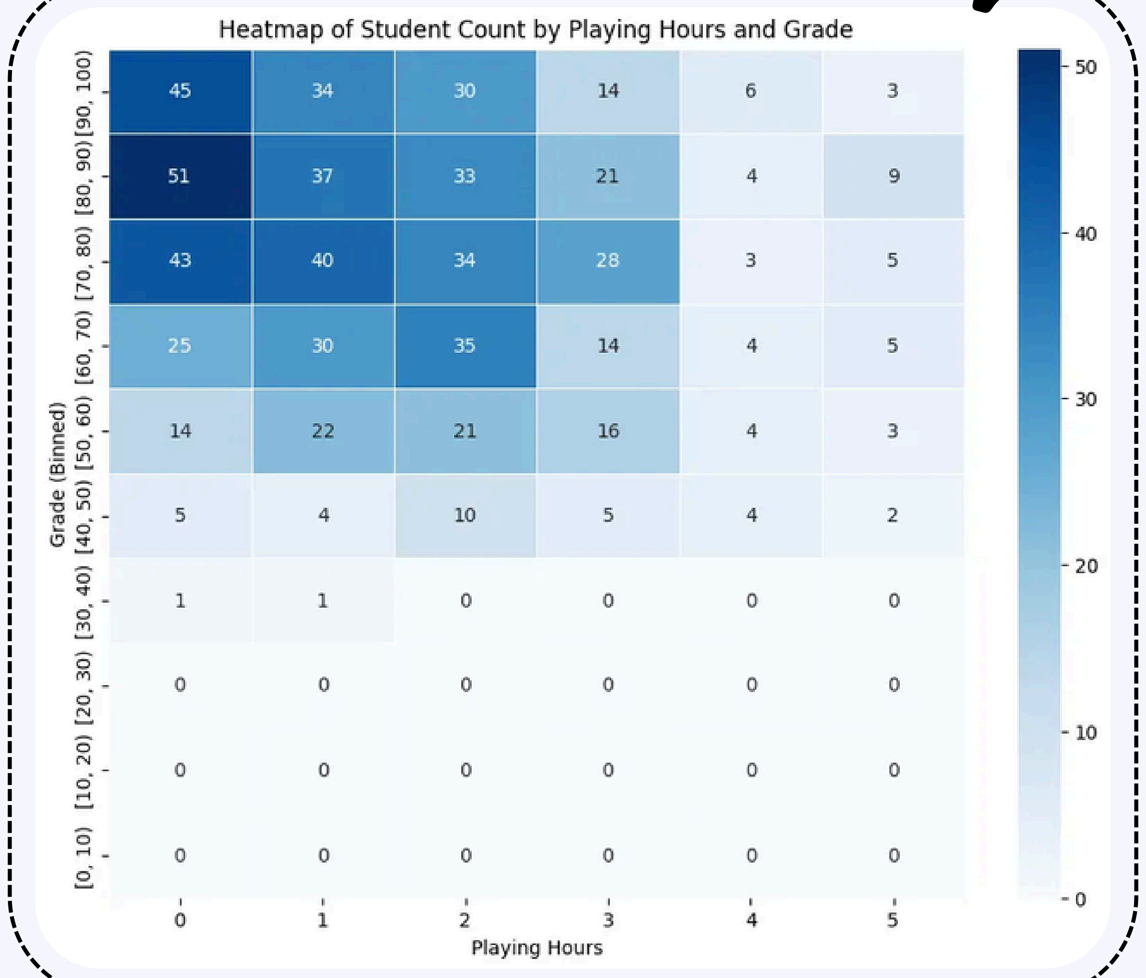
our **X- axis** (Playing hours, Parent Revenue, Parent Education) are **all ordinal** and **discrete data**.

So, it turns out to be different than what we expected (not suitable for clustering).



Alternative way to visualize data :

What if we use **heatmap**



But... This raises a concerns

Is a heatmap a suitable way to visualize relationships in this discrete data instead of using k-means, or is there a better alternative we could try?

Classification

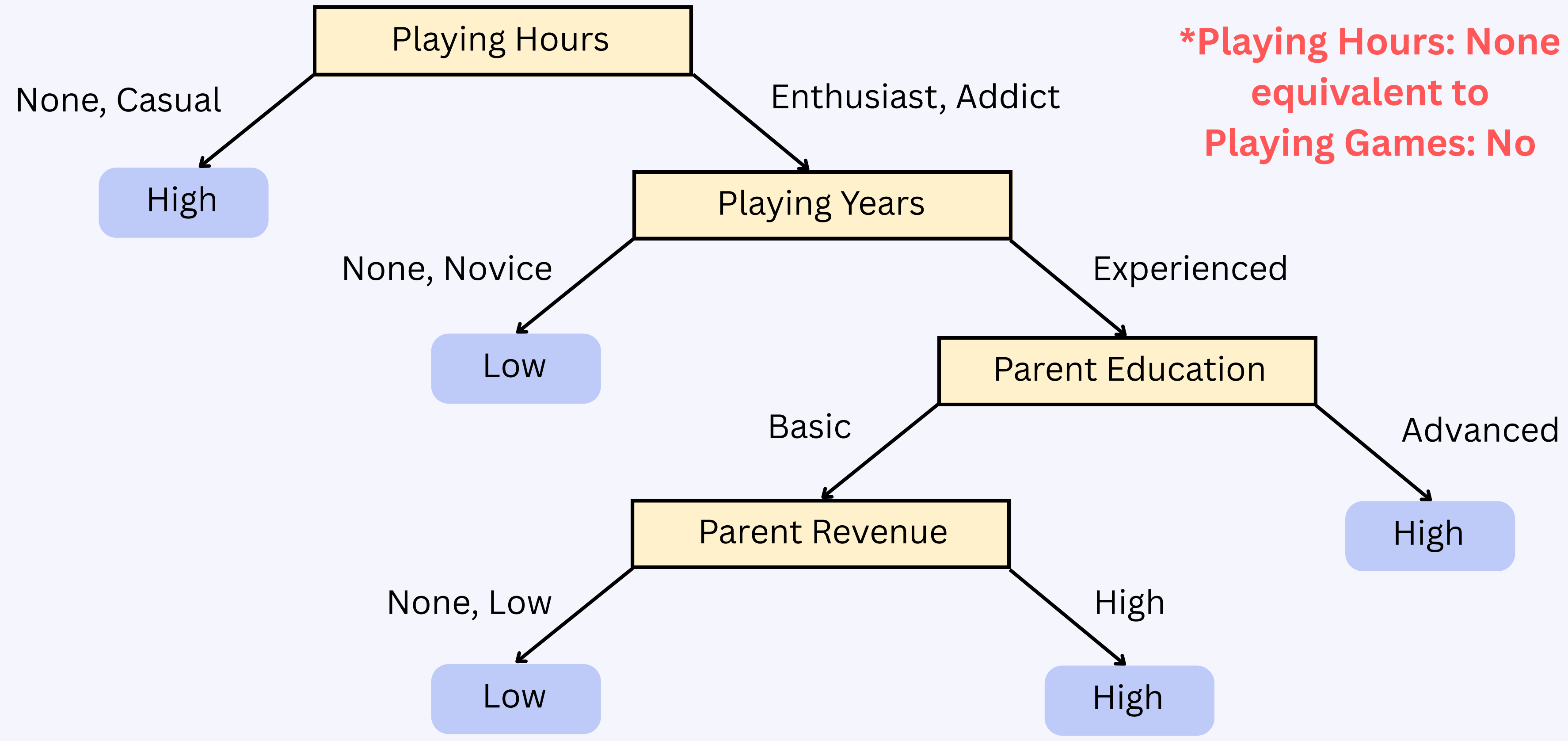
Objectives

- Construct decision tree & classify data into labels (Grade).
- Train simple DNN model to predict label of unseen data.

Does sex matter?

ID	Sex	Playing Years	Playing Hours	Playing Games	Parent Revenue	Parent Education	Grade
1	0	Novice	Casual	Yes	High	Advanced	High
2	1	Novice	Casual	Yes	Low	Basic	High
3	0	None	None	No	Low	Basic	High
4	0	Experienced	Casual	Yes	Low	Basic	Low
5	1	Experienced	Addict	Yes	Low	Advanced	High
6	0	Novice	Casual	Yes	Low	Basic	High
7	1	None	None	No	High	Advanced	???

Possible Decision Tree Deduction

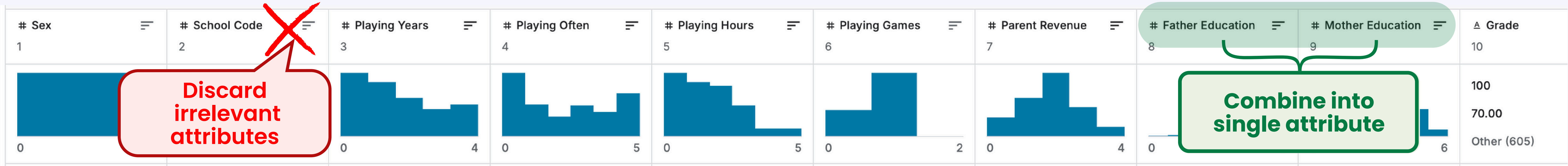


*Playing Hours: None equivalent to Playing Games: No

Finding out which attributes in our dataset matters → Interpreting the data values from dataset → Deciding on the best split & split criterion for decision tree

Classification

Attributes in the Game & Grade Dataset



Interpretation of Attributes for Classification Purposes

	Sex	Playing Years	Playing Hours	Playing Games	Parent Revenue	Father Education	Mother Education	** Parent Education	Grade
	S. Binary	Numerical	Numerical	Binary	Ordinal*	Ordinal*	Ordinal*	Ordinal*	Numerical
Range from Raw Data	0 (F) 1 (M)	0~4	0~5	0/1	0~4	0~6	0~6	0~6	0~100
Range for Decision Tree	0 (F) 1 (M)	0 (None) 1~2 (Novice) 3~4 (Experienced)	0 (None) 1~2 (Casual) 3~4 (Enthusiast) 5 (Addict)	0 (Yes) 1 (No)	0 (None) 1-2 (Low) 3-4 (High)	Combined for Parent Education	Combined for Parent Education	(FE + ME) / 2 ** Equal weights 0.0~3.0 (Basic) 3.5~6.0 (Advanced)	< Median (Low), ≥ Median (High)

Parent Revenue: Which is more optimal?

0	1	2	3	4
None	LOW		HIGH	

Playing Hours

None | Casual   Enthusiast | Addict

Split further??

Use median as threshold for splitting the continuous valued label classes

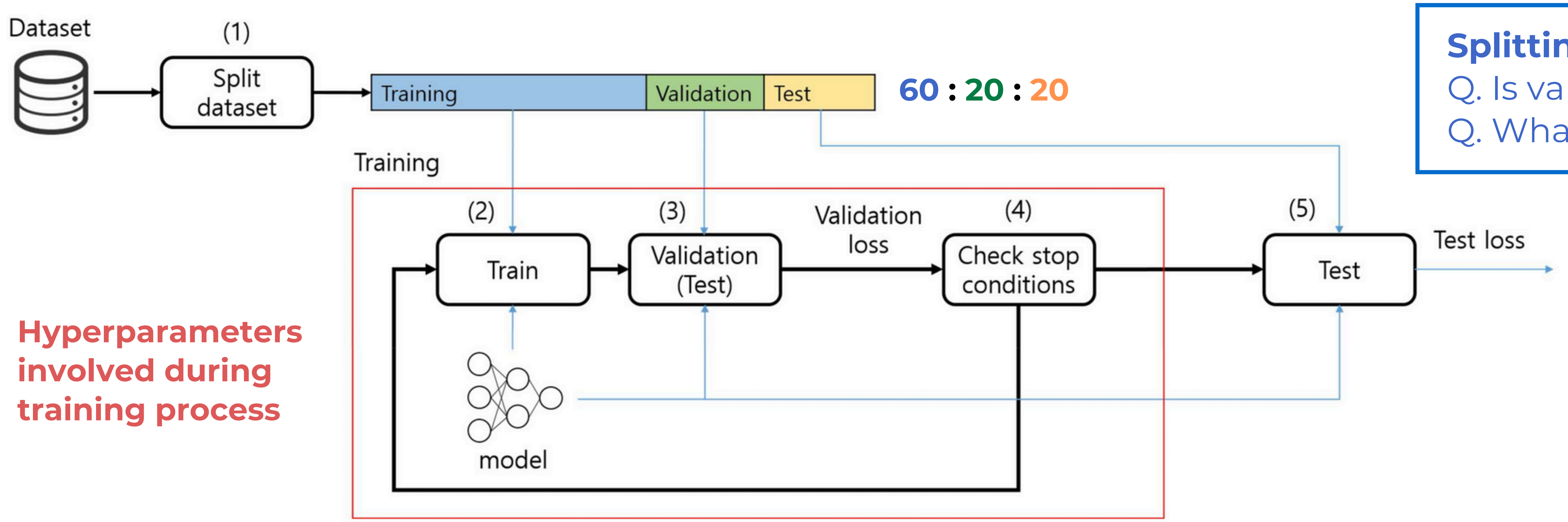
Total rows : 770

Median : $(X_{385} + X_{386}) / 2 = 80.00$

≥ Median (High)

< Median (Low)

Classification



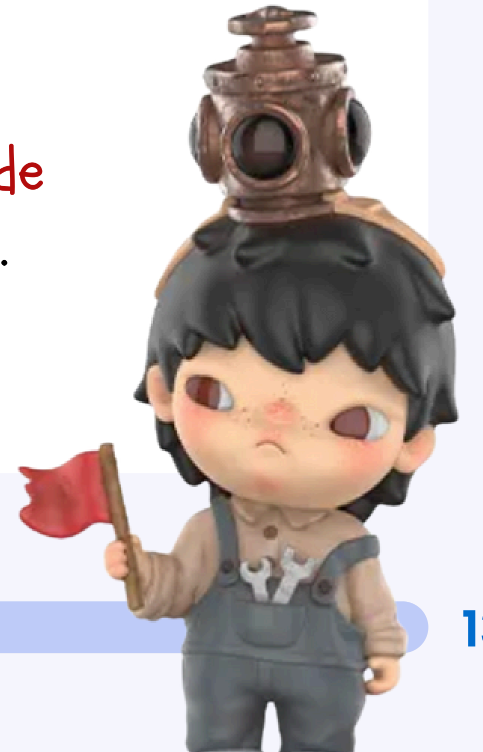
Hyperparameters involved during training process

Splitting dataset into training & test set.
Q. Is validation set necessary in our case?
Q. What split ratio, sampling method?

Measuring split quality/criteria	**GINI Index** **Entropy**
Stopping criteria	**min_samples_split** **min_samples_leaf** **max_depth/max_node_size**

Whats on our mind?

Train model to successfully predict student Grade (High/Low) based on features from our dataset. (gaming behaviours & household factors)



Pattern Mining

What Are We Trying to Find Out?

- Which student characteristics tend to appear together?
- Which behaviors are frequently associated with high or low grades?

e.g. **Does high gaming hour, low parental income → low grades?**

Pattern Mining

e.g. {PlayingHours=High, ParentEdu=Low} ⇒ {Grade=Low}
- 50 out of 770 students match the rule:
  support = 50/770 = 6.5%
- 70 students have {PlayingHours=High, ParentEdu=Low}, and 50 of them have {Grade=Low}:
  confidence= 50/70 ≈ 71%

1. Data Preprocessing
- Continuous data (parent education, grade) and ordinal numerical (ex. playing hours, parent revenue) are binned into categorical data (e.g., High, Low)
- **Convert data into transaction**

Sex	Playing Years	Playing Hours	Parent Revenue	Parent Education	Grade
0	1	1	4	4.5	77.5

{Sex=0, Playing Years=Novice, Playing Hours=Casual, Parent Revenue=High, Parent Education= Advanced, Grade=Low}

4. Interpreting result
- Is the pattern commom? → ***Support***
- Is the rule reliable? → ***Confidence***

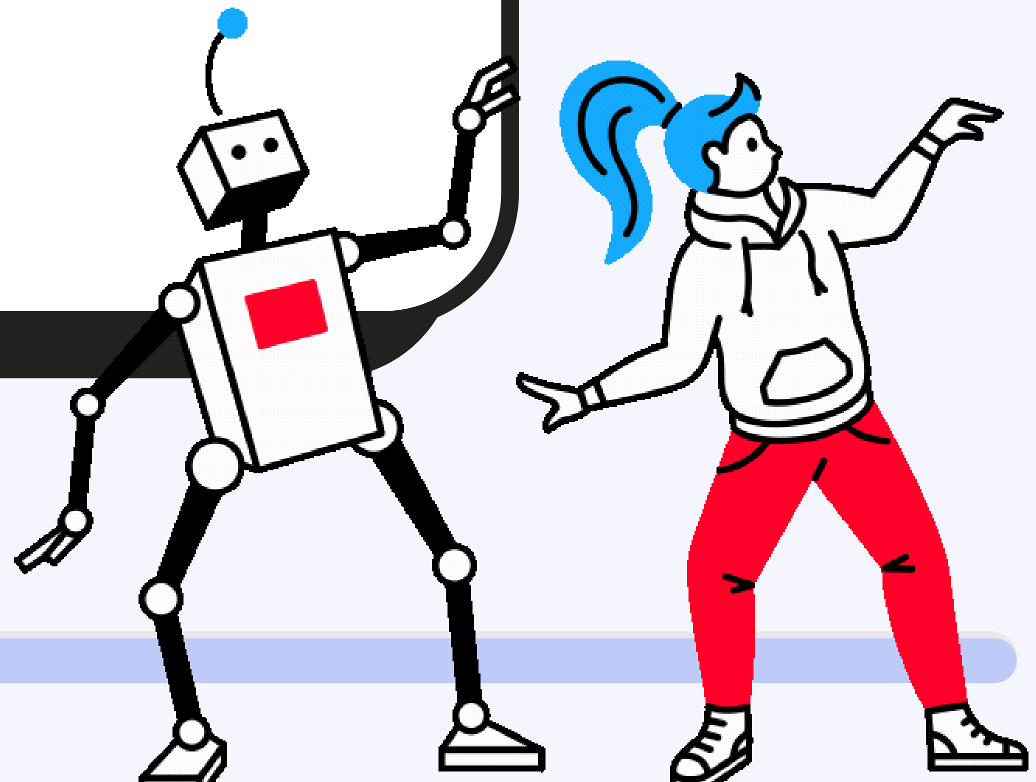
2. Apriori for pattern mining
- Easy to use and interpret
- Works well with our dataset (~770)

3. Support threshold selection
- Initial: min_support = 0.1
- Adjust the support based on observed result:
  many general rules → increase to 0.15
  too few rules → decrease to 0.07

to find pattern in at least 10% of the students

However...

1. Is our minimum support selection strategy sufficient?
2. Are frequent combinations truly related — or just coincidental?



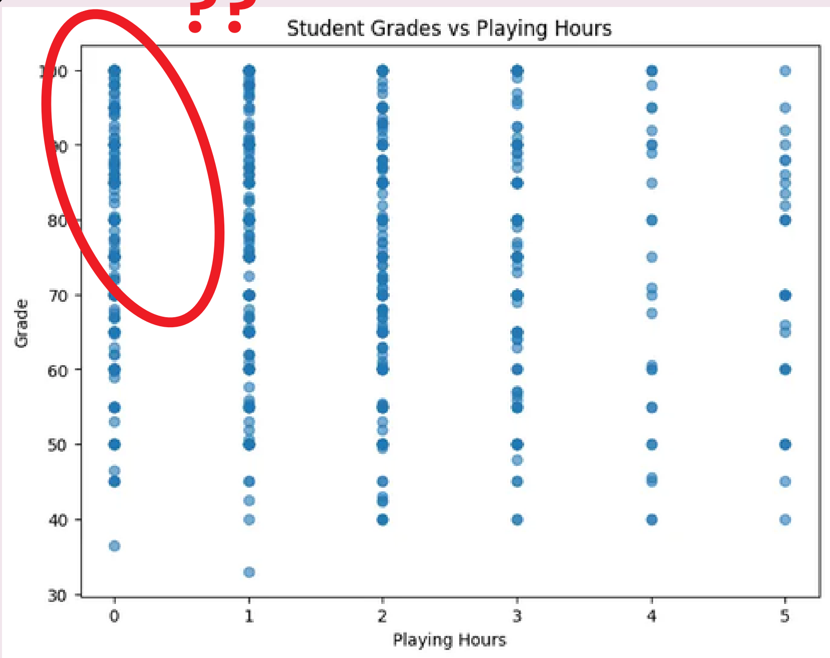
Discussion and Q&A

Clustering helps with revealing underlying structures or patterns that may not be immediately apparent. But in our case,

Is clustering with Discrete Values truly effective?

Clustering with K-Means++

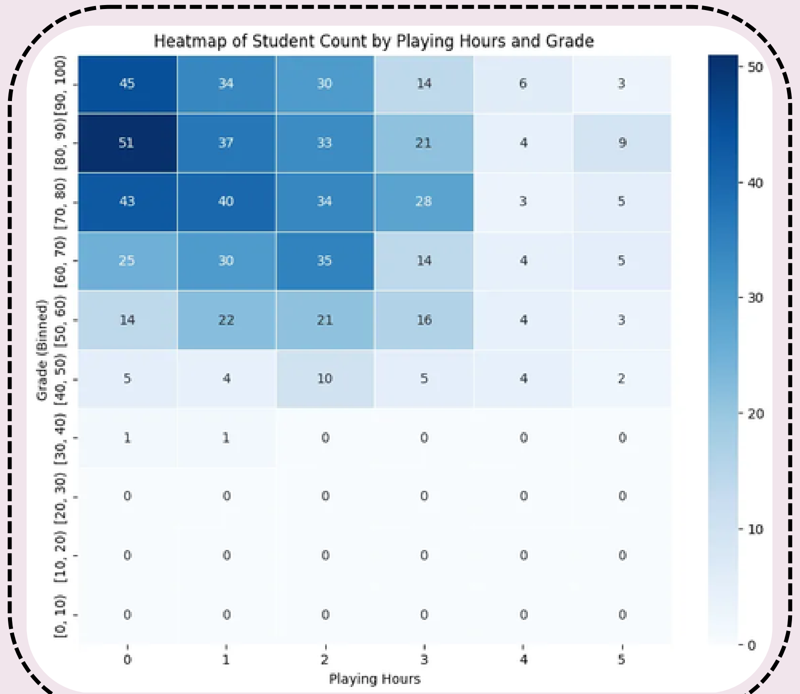
??



Difficult to visually identify clusters due to shape of graph

OR

Heatmap



Shows clear pattern of data density distribution

Challenges with deciding on classification factors

Classification

Parent Revenue: Which is more optimal?

0	1	2	3	4

None	LOW	HIGH

Playing Hours

None | Casual

Enthusiast Addict

Split further??

Use median as threshold for splitting the continuous valued label classes

Total rows : 770

Median : $(X_{385} + X_{386}) / 2 = 80.00$

≥ Median (High)

< Median (Low)

Thank you!