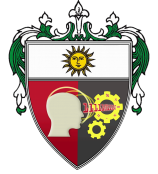




UNIVERSITY OF SANTO TOMAS
College of Information and Computing Sciences
Department of Computer Science



Educational Access in the Philippines:
Identifying Enrollment Disparities Across Regions, Sectors, and Genders

By:

Barte, Marianillete C.

Buhay, Kyle Andrei C.

Chen, Maria Josephine B.

Consorte, Aaryanah Micah T.

Morales, Kyle Gweneth Q.

Instructor:

Adrian Camilo C. Gernale, PSMDS

December 2024

Abstract

The analysis gives a closer look on educational enrollment disparities in the Philippines using the HistoricalEnrollmentData dataset, sourced from the Department of Education and Open Data Philippines. Covering S.Y. 2010-2011 to S.Y. 2020-2021, the dataset provides insights into enrollment patterns across regions, sectors, grade levels, and genders. In spite of progress in education accessibility, significant challenges remain, such as poverty, inadequate infrastructure, and the digital divide, exacerbated during the COVID-19 pandemic.

The disparities in enrollment are identified, with urbanized regions like NCR and Region 4A consistently showing higher enrollment, while rural areas such as CAR and CARAGA lag behind. Machine learning implementations like K-Means clustering and XGBoost predictions highlight patterns and enable predictions for future enrollment trends. Notable findings include a sharp decline in enrollment from 2019-2020 to 2020-2021 and lower enrollment in secondary education, particularly in Grades 9 and 10, pointing to retention challenges.

Overall, the analysis emphasizes the need for targeted interventions, including resource allocation to underserved regions, improving public school infrastructure, and increasing accessibility to state universities and local colleges. The findings aim to inform policymakers and stakeholders in the Philippines to address educational disparities and promote equitable access to education for Filipino students.

Table of Contents

Abstract.....	1
Table of Contents.....	1
Chapter I Introduction.....	1
A. Background of the Dataset.....	1
B. Source of Data and Its Role.....	1
C. Dataset Content.....	2
D. Statement of the Problem.....	3
E. Objectives of the Analysis.....	4
Chapter II Methodology.....	5
A. Data Pipeline.....	5
B. Methods.....	6
B.1. Data Profiling.....	6
B.1.a. Data Type Information.....	7
B.1.b. Data Shape.....	8
B.1.c. Statistical Information.....	8
B.1.d. Comparison of Unique Values.....	9
B.1.e. Checking of Null Values.....	10
B.1.f. Checking of Incorrect Formatting.....	11
B.2. Data Cleaning.....	12
B.2.a. Fixing Incorrect Format.....	13
B.2.b. Replacing Null Values.....	14
B.2.c. Standardizing Columns.....	14
B.2.d. Checking for Outliers.....	15
B.2.e. Checking for Duplicates.....	17
B.3. Feature Engineering.....	17
C. Analysis Process and Techniques Used.....	19
C.1. Descriptive Analysis.....	19
C.2. Exploratory Data Analysis and Visualization.....	20
C.3. Clustering.....	21
C.3.a. Find optimal clusters using Elbow Method.....	21
C.3.b. Fit the KMeans Model and Assign Cluster Labels.....	22
C.3.c. Cluster Summary.....	22
C.4. XGBoost.....	23

C.4.a. Model Training and Testing.....	24
C.4.a Model Application.....	24
Chapter III Results.....	25
A. Visualization and Their Interpretations.....	25
A.1. Total Enrollment by Region (2010-2011 to 2020-2021, in Millions).....	25
A.2. Sector-Wise Enrollment by Region (2010-2011 to 2020-2021).....	26
A.3. Sectoral Distribution.....	27
A.4. School Year Enrollee Trend based on Regions (in Millions).....	28
A.5. Enrollment by Gender and Region (in Millions).....	29
A.6. Total Enrollees by Grade level and Gender (in Millions).....	30
B. Machine Learning Model Implementation and Evaluation.....	33
B.1. K-Means Model.....	33
B.1.a. Elbow Method.....	33
B.1.b. K-Means Clustering.....	34
B.1.c. K-Means Clustering with Principal Component Analysis (PCA).....	36
B.1.d. Cluster Summary.....	37
B.1.e. Comparison of K-Means Clustering.....	42
B.2. XGBoost Model.....	44
B.2.a. Evaluation.....	44
B.2.b. Prediction vs Actual.....	47
B.2.c. Predicted Number of Enrollees in Future School Years.....	47
Chapter IV Recommendations and Conclusions.....	49
A. Summary.....	49
B. Conclusion.....	50
C. Recommendations.....	54
References.....	56
Appendices.....	57
A. Methods: Standardizing Columns.....	57
B. Relevant Code Outputs.....	58
B.1. Dataset Shapes.....	58
B.2. Dataset Information.....	58
B.3. Dataset Describe.....	59

Chapter I Introduction

A. Background of the Dataset

The HistoricalEnrollmentData dataset serves as the foundation of the analysis. It provides a detailed record of educational enrollment trends in the Philippines. This dataset was initially created on September 9, 2024, and had its most recent update on November 22, 2022. It includes comprehensive data spanning multiple academic years, capturing information on the number of enrollees by *region*, *sector*, *grade level*, and *gender*.

The dataset was curated to support educational research and planning, offering insights into disparities and trends across various dimensions. Also, its structured format allows for robust preprocessing, exploration, and statistical modeling. HistoricalEnrollmentData's reliability is enhanced by its standardized form and extensive coverage, making it a good resource for understanding enrollment dynamics and informing evidence-based decision-making in education policy in the Philippines.

B. Source of Data and Its Role

The HistoricalEnrollmentData dataset was obtained from the Department of Education (DepEd) of the Philippines website. DepEd is the primary government agency tasked with managing the country's basic education system. This includes both formal and non-formal learning pathways for Filipinos. Its mandate involves formulating, implementing, and coordinating educational policies to ensure equitable access to quality education for learners (Department of Education, n.d.). Additionally, the dataset is also available in Open Data Philippines (ODPH), an online repository of data from different Philippine government agencies.

This dataset holds a role in monitoring enrollment trends across regions, sectors, and genders in the Philippines. It serves as a tool for policymakers and stakeholders to enable them to evaluate enrollment rates, identify disparities, and implement informed strategies based on statistical evidence. By leveraging insights from this dataset, the authors can develop targeted interventions to uphold DepEd's mission of ensuring that "no Filipino learner is left behind" (Department of Education, n.d.).

In addition to supporting educational planning, the HistoricalEnrollmentData contributes to DepEd's broader objectives of achieving inclusive and equitable education. Of which it is aligned with the Philippines' development goals. The dataset's utility also lies in its ability to highlight areas requiring immediate attention, facilitating data-driven decision-making and possible resource allocation.

C. Dataset Content

HistoricalEnrollmentData provides statistics on historical enrollment figures for both Elementary and Junior High School levels in the Philippines. It spans a decade, covering the academic years from 2010-2011 to 2020-2021. The dataset offers detailed breakdowns across several important dimensions. It also includes data for Public and Private institutions, as well as State Universities and Colleges (SUCs) and Local Universities and Colleges (LUCs). The following is a summary of the columns present in the dataset:

1. Number of Students
2. Region: Indicate the region in the Philippines where the student is enrolled.
3. Gender: Classify the students as either Male or Female.
4. Grade Level: Specify the student's grade level for the school year.

5. School Year: Indicate the academic year in which the student is enrolled (e.g., 2023-2024).
6. Sector: Categorize students based on whether they are from public schools, private schools, or SUCs or LUCs.

D. Statement of the Problem

In spite of prioritizing education since independence in 1946, the Philippine education system still encounters significant challenges that limit access for many Filipinos. These issues are complex and multifaceted, including poverty, armed conflict, inadequate resources and infrastructure, and the digital divide. (Bai, 2023).

Another pressing issue is the digital divide, which has become even more pronounced in the wake of the COVID-19 pandemic. The shift to online learning exposed how many students, particularly in rural and underserved areas, lack access to reliable internet and digital devices, further exacerbating existing educational inequalities. According to the World Bank (2020), the lack of technological access has left many students unable to participate in remote learning, creating a significant barrier to education during the pandemic.

In light of these challenges, this analysis is conducted to identify specific regions, sectors, and grade levels in the country with the lowest enrollment figures. By exploring these areas, the goal is to reveal where access to education is most restricted and highlight disparities in enrollment rates. This thorough examination will pinpoint demographics and geographic locations that are particularly underserved and in need of additional resources and support, which can be a stepping stone for improving educational access and deal with the issues in education.

E. Objectives of the Analysis

The primary objectives of this analysis are the following:

1. Analyze enrollment trends across *genders*, *regions*, *sectors*, and *grade levels* to identify significant disparities and patterns in enrollment rates among various demographic groups.
2. Identify *regions*, *sectors*, or *grade levels* with the lowest enrollment and investigate potential factors contributing to these trends.
3. Utilize KMeans clustering to group regions based on enrollment patterns, including factors such as *enrollment count*, *gender disparity*, *sector type*, and *grade level*, to create targeted interventions based on identified clusters.
4. Use the XGBoost to predict the number of enrollees in the following years. This will provide an additional insight to expect the increase or decrease in enrollment in order to plan accordingly.

Chapter II Methodology

A. Data Pipeline

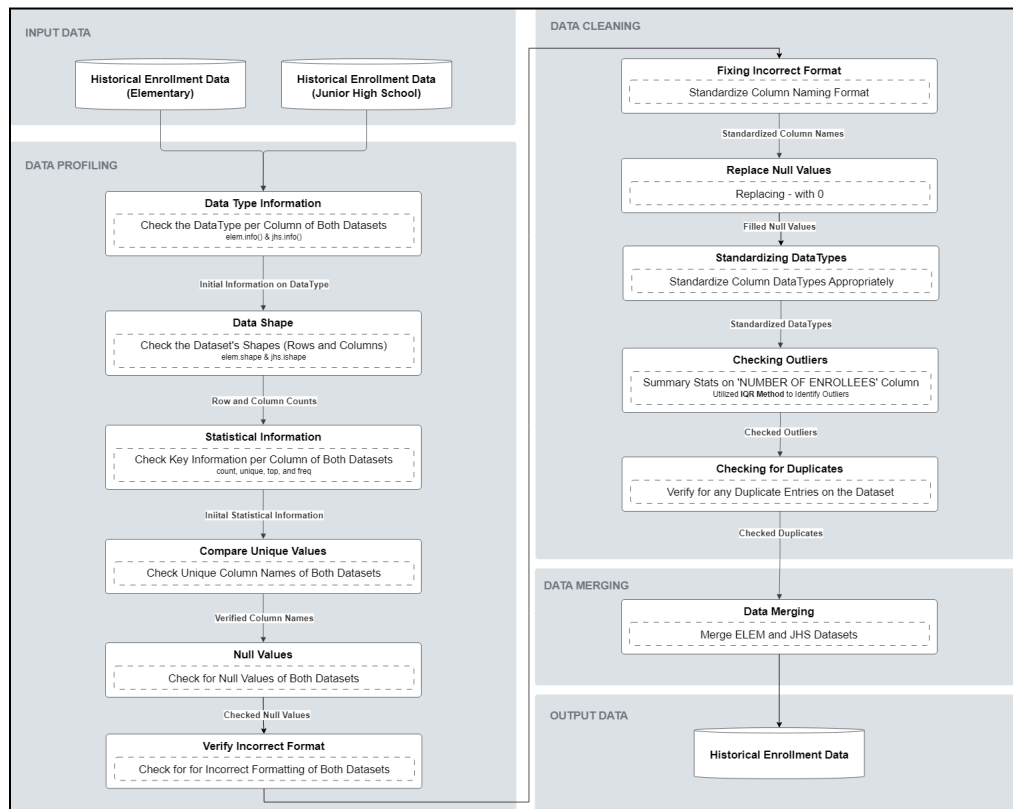


Figure 2.1: Data Pipeline Diagram

Figure 2.1 illustrates the data pipeline diagram to ensure the integrity and consistency of two datasets—Elementary and Junior High School enrollment data—before merging them into a single dataset. This process follows a structured approach to first prepare, clean, validate, and integrate the data, minimizing errors and enhancing reliability for analysis.

First, the pipeline begins with inspecting the structure and summary of both datasets. The information method provides details on data types, null values, and memory usage. Whereas the shape and describe functions summarize the dataset dimensions and key statistics. Then, each column in both datasets is validated to ensure consistency to expected formats. ‘GRADE

LEVEL,' 'REGION,' 'GENDER,' 'SECTOR,' and 'SCHOOL YEAR' are checked for unique values to identify inconsistencies. Null values are identified and corresponding records are further checked and inspected for potential corrections or exclusions also.

Afterwards, data cleaning begins with white space removal by stripping unnecessary spaces in categorical fields such as 'REGION' and 'GENDER.' In relation to numerical formatting, commas and spaces are removed in the 'NUMBER OF ENROLLEES' column. Also, dashes are replaced with zeros. Converted columns such as 'SCHOOL YEAR,' 'SECTOR,' 'GRADE LEVEL,' 'REGION,' and 'GENDER' are type converted to categorical types accordingly.

Subsequently, outliers were identified and handled using the Interquartile Range (IQR) method. Also, duplicate records were removed, retaining the first instance to ensure consistency throughout the HistoricalEnrollmentData. Lastly, 'NON-GRADE' values were reclassified as 'NON-GRADE (PRIMARY)' for elementary and 'NON-GRADE (SECONDARY)' for Junior High School dataset to differentiate the levels.

Lastly, the cleaned datasets were merged into a single dataset using common columns such as 'SCHOOL YEAR,' 'SECTOR,' 'REGION,' 'GRADE LEVEL,' and 'GENDER.' The authors' employed an outer join to ensure that no data was lost during the merge process. The final merged dataset was exported to a CSV file for further analysis.

B. Methods

B.1. Data Profiling

Before proceeding with analysis, the data first undergo profiling, which seeks to know details such as the data type, shape, count, format, and values. The results of the data profiling are as follows:

B.1.a. Data Type Information

Based on the conducted operation to get the data types, it was revealed that all columns under the elementary dataset and junior high school dataset are all stored as object type. Tables 2.1 and 2.2 show information for both elementary and junior high school dataset.

Table 2.1: Data Information: Elementary

Column	Non-Null Count	Data Type
School Year	8,568 <i>non-null</i>	object
Sector	8,568 <i>non-null</i>	object
Region	8,568 <i>non-null</i>	object
Grade Level	8,568 <i>non-null</i>	object
Gender	8,568 <i>non-null</i>	object
Number of Enrollees	8,568 <i>non-null</i>	object
Memory Usage		401.8 kB

Object types may include mixed data types (combination of text, numbers, etc.) and non-standard data. This information about the data type implies the need for type-conversion in the following steps for better handling of values, specifically for features that are required to be numerical data.

Table 2.2: Data Information: Junior High School

Column	Non-Null Count	Data Type
School Year	5,202 <i>non-null</i>	object
Sector	5,202 <i>non-null</i>	object
Region	5,202 <i>non-null</i>	object
Grade Level	5,202 <i>non-null</i>	object

Gender	5,202 <i>non-null</i>	object
Number of Enrollees	5,202 <i>non-null</i>	object
Memory Usage		244.0 kB

B.1.b. Data Shape

The elementary dataset contains 8,568 rows (records) and 6 columns (features), while the junior high school dataset consists of 5,202 rows with the same number of columns. Both datasets include details on *region*, *sector*, *grade level*, *gender*, *school year*, and *enrollment count*.

Table 2.3: Data Shape: Elementary and Junior High School

	Rows	Columns
Elementary	8,568	6
Junior High School	5,202	6

B.1.c. Statistical Information

The statistical information on both datasets, shown in Tables 2.4 and 2.5, are also gathered specifically the total number of entries in each column ('count'), the number of unique values ('unique'), most frequent value ('top'), and the number of times the most frequent value appears ('freq').

The analysis of 'freq' and 'top' indicates that public schools and Region 1 have the highest representation, pointing to regional and sector-specific trends. Male enrollment is more prevalent in elementary, while female enrollment becomes dominant in junior high school, reflecting potential analysis for gender-based patterns. Enrollment is particularly high in certain grade levels, such as Kindergarten and Grade 7, likely due to population

surges or targeted programs. Furthermore, the dataset is deemed reliable for analysis, as the 'count' shows consistent row numbers across all columns.

Table 2.4: Statistical Information: Elementary

	School Year	Sector	Region	Grade Level	Gender	No. of Enrollees
Count	8,568	8,568	8,568	8,568	8,568	8,568
Unique	11	3	34	8	2	5,202
Top	2014-2015	PUBLIC	REGION 1	KINDERG-ARTEN	MALE	-
Frequency	816	2,856	456	1,122	5,284	564

Table 2.5: Statistical Information: Junior High School

	School Year	Sector	Region	Grade Level	Gender	No. of Enrollees
Count	5,202	5,202	5,202	5,202	5,202	5,202
Unique	11	3	17	5	3	3,690
Top	2014-2015	PUBLIC	REGION 1	GRADE 7	FEMALE	-
Frequency	510	1,734	306	1,122	2,601	259

B.1.d. Comparison of Unique Values

There are several differences and similarities between the elementary and junior high school datasets. For 'GRADE LEVEL,' elementary covers grades Kindergarten to grade 6, whereas junior high school covers grades 7 through 10. Both contain a 'NON-GRADE' category. In terms of 'REGION,' both datasets include data from all 17 regions in the Philippines. However, the elementary dataset has some regions with extra

spaces. Similarly, the ‘GENDER’ values under the junior high school dataset also have spacing inconsistencies. As for the ‘SECTOR’ and ‘SCHOOL YEAR’ sections, both datasets hold the same values. These inconsistencies observed in different areas of the datasets suggest a need for data cleaning.

```
def unique_values(column):
    print(f"UNIQUE {column.upper()}")
    print(f"JHS: {jhs[column].unique().tolist()}")
    print(f"ELEM: {elem[column].unique().tolist()}")

columns = ['GRADE LEVEL', 'REGION', 'GENDER', 'SECTOR', 'SCHOOL YEAR']

for column in columns:
    unique_values(column)

UNIQUE GRADE LEVEL
JHS: ['GRADE 7', 'GRADE 8', 'GRADE 9', 'GRADE 10', 'NON-GRADE']
ELEM: ['KINDERGARTEN', 'GRADE 1', 'GRADE 2', 'GRADE 3', 'GRADE 4', 'GRADE 5', 'GRADE 6', 'NON-GRADE']

UNIQUE REGION
JHS: ['REGION 1', 'REGION 2', 'REGION 3', 'REGION 4A', 'REGION 4B', 'REGION 5', 'REGION 6', 'REGION 7', 'REGION 8', 'REGION 9', 'REGION 10', 'REGION 11', 'REGION 12', 'CARAG A', 'BARMM', 'CAR', 'NCR']
ELEM: ['REGION 1', 'REGION 2', 'REGION 3', 'REGION 4A', 'REGION 4B', 'REGION 5', 'REGION 6', 'REGION 7', 'REGION 8', 'REGION 9', 'REGION 10', 'REGION 11', 'REGION 12', 'CARAG A', 'BARMM', 'CAR', 'NCR', 'REGION 1', 'REGION 2', 'REGION 3', 'REGION 4A', 'REGION 4B', 'REGION 5', 'REGION 6', 'REGION 7', 'REGION 8', 'REGION 9', 'REGION 10', 'REGION 11', 'REGION 12', 'CARAG A', 'BARMM', 'CAR', 'NCR']

UNIQUE GENDER
JHS: ['MALE', 'FEMALE', 'MALE']
ELEM: ['MALE', 'FEMALE']

UNIQUE SECTOR
JHS: ['PUBLIC', 'PRIVATE', 'SUSLUKS']
ELEM: ['PUBLIC', 'PRIVATE', 'SUSLUKS']

UNIQUE SCHOOL YEAR
JHS: ['2010-2011', '2011-2012', '2012-2013', '2013-2014', '2014-2015', '2015-2016', '2016-2017', '2017-2018', '2018-2019', '2019-2020', '2020-2021']
ELEM: ['2010-2011', '2011-2012', '2012-2013', '2013-2014', '2014-2015', '2015-2016', '2016-2017', '2017-2018', '2018-2019', '2019-2020', '2020-2021']
```

Figure 2.2: Code Snippet and Output of Unique Values

B.1.e. Checking of Null Values

To validate the datasets, their values are checked if they contain null or missing values. This step revealed that there are no missing values across all columns in both datasets, which is a good start before proceeding further.

```

print('ELEMENTARY')
for column in elem.columns:
    missing_values = elem[elem[column].isna()]
    if not missing_values.empty:
        print(f"Missing values found in column: {column}")
        print(missing_values.head(10))
    else:
        print(f"No missing values in column: {column}")

print('\nJUNIOR HIGH SCHOOL')
for column in jhs.columns:
    missing_values = jhs[jhs[column].isna()]
    if not missing_values.empty:
        print(f"Missing values found in column: {column}")
        print(missing_values.head(10))
    else:
        print(f"No missing values in column: {column}")

```

ELEMENTARY
 No missing values in column: SCHOOL YEAR
 No missing values in column: SECTOR
 No missing values in column: REGION
 No missing values in column: GRADE LEVEL
 No missing values in column: GENDER
 No missing values in column: NUMBER OF ENROLLEES

 JUNIOR HIGH SCHOOL
 No missing values in column: SCHOOL YEAR
 No missing values in column: SECTOR
 No missing values in column: REGION
 No missing values in column: GRADE LEVEL
 No missing values in column: GENDER
 No missing values in column: NUMBER OF ENROLLEES

Figure 2.3: Checking of Null Values

B.1.f. Checking of Incorrect Formatting

The dataset is examined for inconsistencies in formatting to keep the integrity of the data. No issues were found in the columns ‘GRADE LEVEL,’ ‘SCHOOL YEAR,’ and ‘SECTOR.’ However, minimal formatting issues were found in the remaining columns. In ‘REGION’ and ‘GENDER,’ spacing issues were found in one of the two datasets. As for the ‘NUMBER OF ENROLLEES,’ both datasets contained non-numeric characters such as ‘-,’ ‘,’ and ‘.’.

```

Column: SCHOOL_YEAR
[14]:
print('ELEMENTARY')
inc_sy = elem['SCHOOL_YEAR'].str.match(r'^[0-9]{4}$')
if inc_sy.empty():
    print('Incorrect SCHOOL_YEAR formatting: None')
else:
    print('Incorrect SCHOOL_YEAR formatting:')
    print(inc_sy['SCHOOL_YEAR'].unique().tolist())

print('\nJUNIOR HIGH SCHOOL')
inc_sy2 = jhs['SCHOOL_YEAR'].str.match(r'^[0-9]{4}$')
if inc_sy2.empty():
    print('Incorrect SCHOOL_YEAR formatting: None')
else:
    print('Incorrect SCHOOL_YEAR formatting:')
    print(inc_sy2['SCHOOL_YEAR'].unique().tolist())

ELEMENTARY
Incorrect SCHOOL_YEAR formatting: None

JUNIOR HIGH SCHOOL
Incorrect SCHOOL_YEAR formatting: None

Column: GRADE_LEVEL
[15]:
print('ELEMENTARY')
inc_grade = elem['GRADE_LEVEL'].str.match(r'^[0-9]{1}$')
if inc_grade.empty():
    print('Incorrect GRADE_LEVEL formatting: None')
else:
    print('Incorrect GRADE_LEVEL formatting:')
    print(inc_grade['GRADE_LEVEL'].unique().tolist())

print('\nJUNIOR HIGH SCHOOL')
inc_grade2 = jhs['GRADE_LEVEL'].str.match(r'^[0-9]{1}$')
if inc_grade2.empty():
    print('Incorrect GRADE_LEVEL formatting: None')
else:
    print('Incorrect GRADE_LEVEL formatting:')
    print(inc_grade2['GRADE_LEVEL'].unique().tolist())

ELEMENTARY
Incorrect GRADE_LEVEL formatting: None

JUNIOR HIGH SCHOOL
Incorrect GRADE_LEVEL formatting: None

Column: NUMBER OF ENROLLEES
[16]:
print('ELEMENTARY')
inc_num = elem['NUMBER OF ENROLLEES'].apply(lambda x: x.isdigit())
if inc_num.any():
    invalid_values = elem.loc[inc_num, 'NUMBER OF ENROLLEES']
    non_numeric_chars = set(''.join(invalid_values)) - set('0123456789')
    print('Characters found that are not numeric:', non_numeric_chars)
else:
    print('All values are in correct format')

print('\nJUNIOR HIGH SCHOOL')
inc_num2 = jhs['NUMBER OF ENROLLEES'].apply(lambda x: x.isdigit())
if inc_num2.any():
    invalid_values2 = jhs.loc[inc_num2, 'NUMBER OF ENROLLEES']
    non_numeric_chars2 = set(''.join(invalid_values2)) - set('0123456789')
    print('Characters found that are not numeric:', non_numeric_chars2)
else:
    print('All values are in correct format')

ELEMENTARY
Incorrect REGION formatting:
[ ' REGION 1 ', ' REGION 2 ', ' REGION 3 ', ' REGION 4A ', ' REGION 4B ', ' RE
GION 5 ', ' REGION 6 ', ' REGION 7 ', ' REGION 8 ', ' REGION 9 ', ' REGION 10
 ', ' REGION 11 ', ' REGION 12 ', ' CARAGA ', ' BARMM ', ' CAR ', ' NCR ' ]

JUNIOR HIGH SCHOOL
Incorrect REGION formatting: None

```

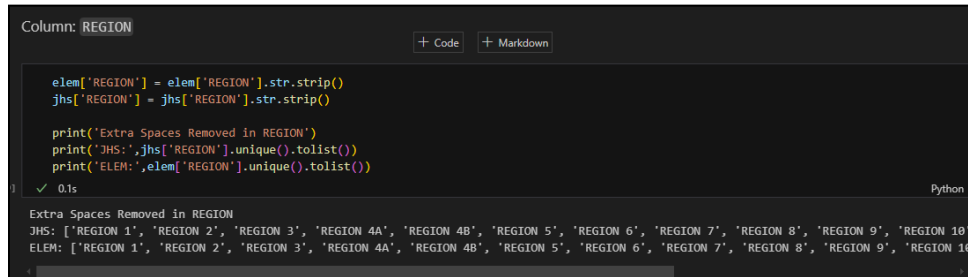
Figure 2.4: Checking for Incorrect Formatting of all Features

B.2. Data Cleaning

After reviewing the dataset's profile, the next step is cleaning the data. Data cleaning is the process of rectifying inconsistencies, errors or inaccuracies in the dataset to improve its quality and ensure it is suitable for analysis. This step involve numerous cleaning tasks such as:

B.2.a. Fixing Incorrect Format

During data profiling, inconsistencies were identified in the 'REGION' variable, such as extra spaces. Additionally, extra spaces and commas were present in the 'NUMBER OF ENROLLEES' column. As a result, the formatting was corrected by removing spaces and the special characters.



```
Column: REGION

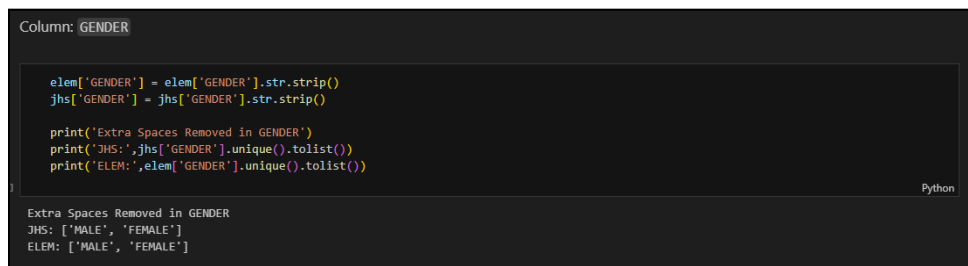
elem["REGION"] = elem["REGION"].str.strip()
jhs["REGION"] = jhs["REGION"].str.strip()

print('Extra Spaces Removed in REGION')
print('JHS:', jhs["REGION"].unique().tolist())
print('ELEM:', elem["REGION"].unique().tolist())
```

✓ 0.1s

Extra Spaces Removed in REGION
JHS: ['REGION 1', 'REGION 2', 'REGION 3', 'REGION 4A', 'REGION 4B', 'REGION 5', 'REGION 6', 'REGION 7', 'REGION 8', 'REGION 9', 'REGION 10']
ELEM: ['REGION 1', 'REGION 2', 'REGION 3', 'REGION 4A', 'REGION 4B', 'REGION 5', 'REGION 6', 'REGION 7', 'REGION 8', 'REGION 9', 'REGION 10']

Figure 2.5: Fixing Incorrect Format: Region



```
Column: GENDER

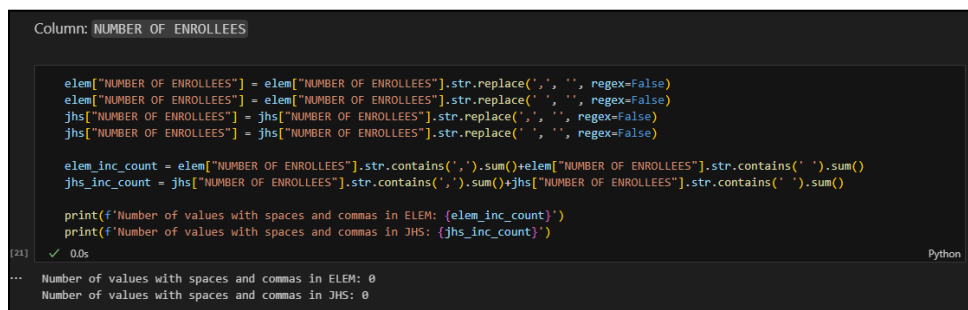
elem["GENDER"] = elem["GENDER"].str.strip()
jhs["GENDER"] = jhs["GENDER"].str.strip()

print('Extra Spaces Removed in GENDER')
print('JHS:', jhs["GENDER"].unique().tolist())
print('ELEM:', elem["GENDER"].unique().tolist())
```

✓ 0.1s

Extra Spaces Removed in GENDER
JHS: ['MALE', 'FEMALE']
ELEM: ['MALE', 'FEMALE']

Figure 2.6: Fixing Incorrect Format: Gender



```
Column: NUMBER OF ENROLLEES

elem["NUMBER OF ENROLLEES"] = elem["NUMBER OF ENROLLEES"].str.replace(' ', '', regex=False)
elem["NUMBER OF ENROLLEES"] = elem["NUMBER OF ENROLLEES"].str.replace(',', '', regex=False)
jhs["NUMBER OF ENROLLEES"] = jhs["NUMBER OF ENROLLEES"].str.replace(' ', '', regex=False)
jhs["NUMBER OF ENROLLEES"] = jhs["NUMBER OF ENROLLEES"].str.replace(',', '', regex=False)

elem_inc_count = elem["NUMBER OF ENROLLEES"].str.contains(' ').sum() + elem["NUMBER OF ENROLLEES"].str.contains(',').sum()
jhs_inc_count = jhs["NUMBER OF ENROLLEES"].str.contains(' ').sum() + jhs["NUMBER OF ENROLLEES"].str.contains(',').sum()

print(f'Number of values with spaces and commas in ELEM: {elem_inc_count}')
print(f'Number of values with spaces and commas in JHS: {jhs_inc_count}')
```

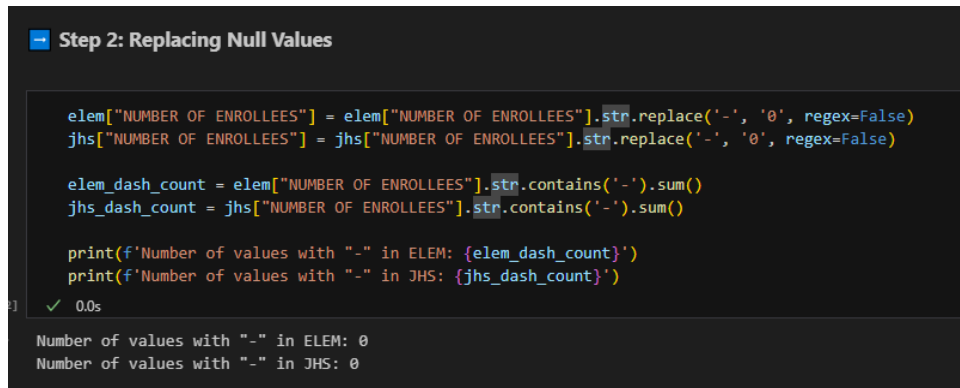
✓ 0.0s

Number of values with spaces and commas in ELEM: 0
Number of values with spaces and commas in JHS: 0

Figure 2.7: Fixing Incorrect Format: Number of Enrollees

B.2.b. Replacing Null Values

The dashes found in the 'NUMBER OF ENROLLEES' variable were interpreted as null values and were replaced with 0, as a dash typically represents the absence of data.



```
Step 2: Replacing Null Values

elem["NUMBER OF ENROLLEES"] = elem["NUMBER OF ENROLLEES"].str.replace('-', '0', regex=False)
jhs["NUMBER OF ENROLLEES"] = jhs["NUMBER OF ENROLLEES"].str.replace('-', '0', regex=False)

elem_dash_count = elem["NUMBER OF ENROLLEES"].str.contains('-').sum()
jhs_dash_count = jhs["NUMBER OF ENROLLEES"].str.contains('-').sum()

print(f'Number of values with "-" in ELEM: {elem_dash_count}')
print(f'Number of values with "-" in JHS: {jhs_dash_count}')
```

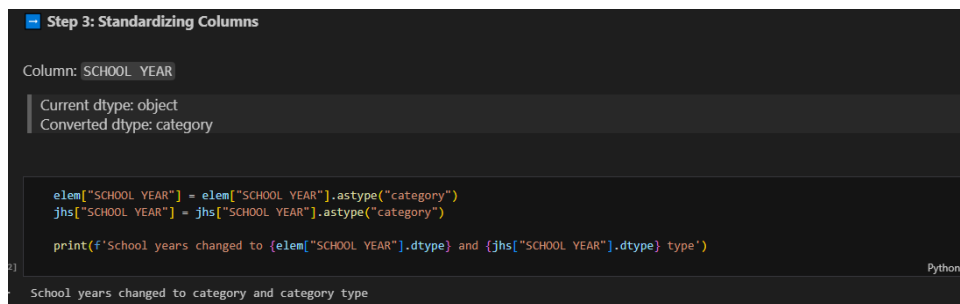
✓ 0.0s

Number of values with "-" in ELEM: 0
Number of values with "-" in JHS: 0

Figure 2.8: Replacing Null Values with 0

B.2.c. Standardizing Columns

During data profiling, incorrect data types were identified in the variables. To address this issue, the columns were standardized by converting the data types of 'SCHOOL YEAR,' 'SECTOR,' 'REGION,' 'GRADE LEVEL,' and 'GENDER' from object to category, while 'NUMBER OF ENROLLEES' was converted from object to int32. Figures 2.9 to 2.11 illustrate the standardization processes on 'SCHOOL YEAR,' 'SECTOR,' and 'NUMBER OF ENROLLEES.' Screenshots of other columns are shown in Appendix A.



```
Step 3: Standardizing Columns

Column: SCHOOL YEAR

Current dtype: object
Converted dtype: category

elem["SCHOOL YEAR"] = elem["SCHOOL YEAR"].astype("category")
jhs["SCHOOL YEAR"] = jhs["SCHOOL YEAR"].astype("category")

print(f'School years changed to {elem["SCHOOL YEAR"].dtype} and {jhs["SCHOOL YEAR"].dtype} type')
```

Python

School years changed to category and category type

Figure 2.9: Standardizing Columns: School Year



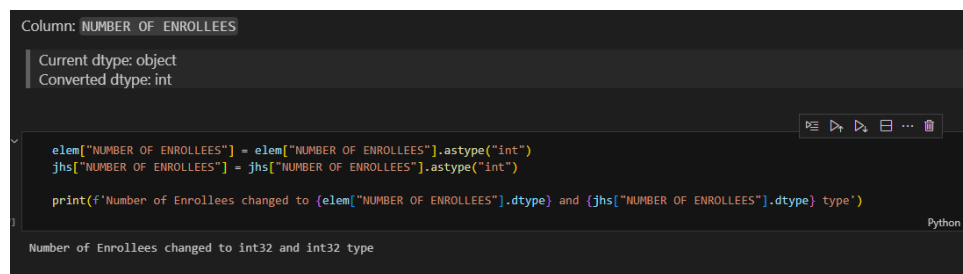
```
Column: SECTOR
Current dtype: object
Converted dtype: category

elem["SECTOR"] = elem["SECTOR"].astype("category")
jhs["SECTOR"] = jhs["SECTOR"].astype("category")

print(f'Sectors changed to {elem["SECTOR"].dtype} and {jhs["SECTOR"].dtype} type')
```

Sectors changed to category and category type

Figure 2.10: Standardizing Columns: Sector



```
Column: NUMBER OF ENROLLEES
Current dtype: object
Converted dtype: int

elem["NUMBER OF ENROLLEES"] = elem["NUMBER OF ENROLLEES"].astype("int")
jhs["NUMBER OF ENROLLEES"] = jhs["NUMBER OF ENROLLEES"].astype("int")

print(f'Number of Enrollees changed to {elem["NUMBER OF ENROLLEES"].dtype} and {jhs["NUMBER OF ENROLLEES"].dtype} type')
```

Number of Enrollees changed to int32 and int32 type

Figure 2.11: Standardizing Columns: Number of Enrollees

B.2.d. Checking for Outliers

To be able to distinguish if there are outliers within the dataset, the authors analyze the summary statistics of the ‘NUMBER OF ENROLLEES’ column for each dataset. This will help determine the range of the data.

After describing the summary of statistics, shown in Tables 2.6 and 2.7, the authors’ use the Interquartile Range (IQR) Method to determine outliers. With that, In both of the datasets the maximum values far surpass the thresholds determined by the IQR, indicating potential outliers. Additionally, the large standard deviation and uneven data distribution further highlight the presence of extreme outliers. Given these, we decided to retain the outliers as they could provide insights for the analysis.

Table 2.6: Description of Number of Enrollees in the Elementary Dataset

Summary Statistics: Elementary Dataset	
Count	8,568.00
Mean	20,578.69
Standard Deviation	31,830.63
Minimum	0.00
25% Percentile	64.00
50% Percentile	2,392.50
75% Percentile	35,727.25
Maximum	178,471.00

Table 2.7: Description of Number of Enrollees in the JHS Dataset

Summary Statistics: Junior High School Dataset	
Count	5202.00
Mean	1,6061.31
Standard Deviation	24,028.50
Minimum	0.00
25% Percentile	235.25
50% Percentile	4,237.50
75% Percentile	25,188.00
Maximum	139,254.00

B.2.e. Checking for Duplicates

```
elem_duplicates = elem[elem.duplicated(keep=False)].sort_values(by='NUMBER OF ENROLLEES', ascending=False)
jhs_duplicates = jhs[jhs.duplicated(keep=False)].sort_values(by='NUMBER OF ENROLLEES', ascending=False)

print("Duplicates in Elementary DataFrame (sorted):")
print(elem_duplicates)

print("\nDuplicates in JHS DataFrame (sorted):")
print(jhs_duplicates)

elem = elem.drop_duplicates(keep='first')
jhs = jhs.drop_duplicates(keep='first')

elem_duplicates = elem[elem.duplicated(keep=False)].sort_values(by='NUMBER OF ENROLLEES', ascending=False)
jhs_duplicates = jhs[jhs.duplicated(keep=False)].sort_values(by='NUMBER OF ENROLLEES', ascending=False)

print("Duplicates in Elementary DataFrame after Dropping (sorted):")
print(elem_duplicates)
print("\nDuplicates in JHS DataFrame after Dropping (sorted):")
print(jhs_duplicates)
```

[35] ✓ 0.0s

```
... Duplicates in Elementary DataFrame (sorted):
Empty DataFrame
Columns: [SCHOOL YEAR, SECTOR, REGION, GRADE LEVEL, GENDER, NUMBER OF ENROLLEES]
Index: []

Duplicates in JHS DataFrame (sorted):
Empty DataFrame
Columns: [SCHOOL YEAR, SECTOR, REGION, GRADE LEVEL, GENDER, NUMBER OF ENROLLEES]
Index: []

Duplicates in Elementary DataFrame after Dropping (sorted):
Empty DataFrame
Columns: [SCHOOL YEAR, SECTOR, REGION, GRADE LEVEL, GENDER, NUMBER OF ENROLLEES]
Index: []

Duplicates in JHS DataFrame after Dropping (sorted):
Empty DataFrame
Columns: [SCHOOL YEAR, SECTOR, REGION, GRADE LEVEL, GENDER, NUMBER OF ENROLLEES]
Index: []
```

Figure 2.12: Checking for duplicates

Figure 2.12 depicts code that finds and displays all the duplicate rows through the ‘NUMBER OF ENROLLEES.’ Afterwards, It removes duplicates of each duplicated entry then checks it for anything that is remaining for each of the data frames in elementary and junior high school. Thus, it resulted with no duplicates.

B.3. Feature Engineering

B.3.a. Sector Enrollment Percentage

The *Sector Enrollment Percentage* feature is the percentage of total enrollment within a specific sector. It provides insights into how enrollment is distributed across different sectors within each region.

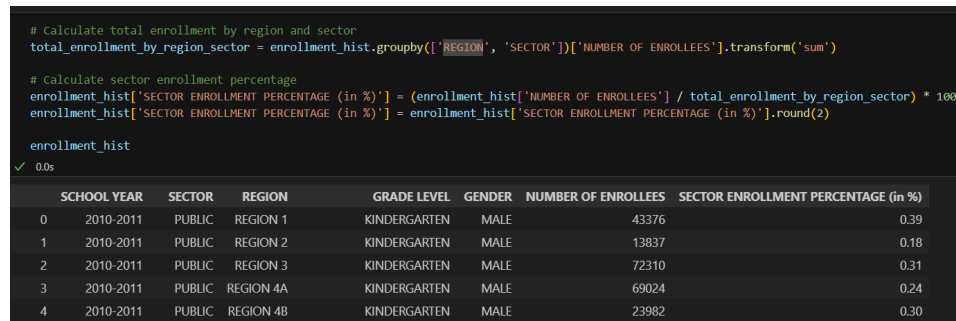


Figure 2.13: Sector Enrollment Percentage (in %) Column

B.3.b. Average Class Size

The *Average Class Size* feature calculates the average enrollment per ‘GRADE LEVEL’ in each ‘REGION’, revealing enrollment distribution patterns. This provides insights into how crowded or evenly distributed the enrollment is across grade levels within regions in the Philippines.

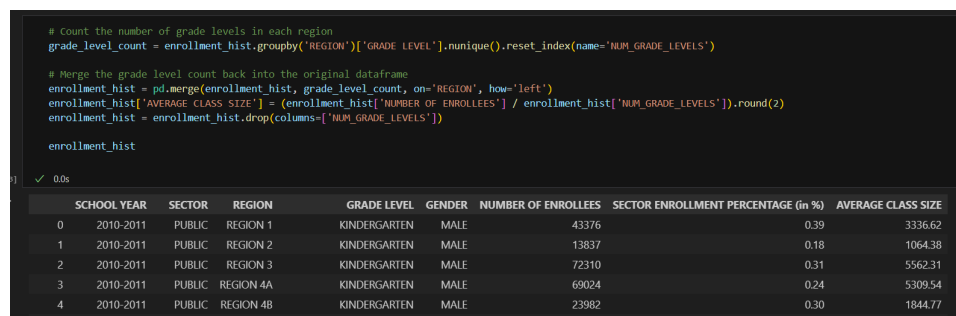


Figure 2.14: Average Class Size Column

B.3.c. Growth Rate

The *Growth Rate* feature presents a yearly percentage change in enrollment given the ‘REGION,’ ‘SECTOR,’ and ‘GRADE LEVEL.’ It provides how enrollment trends evolve over time and shows its rapid growth and decline.

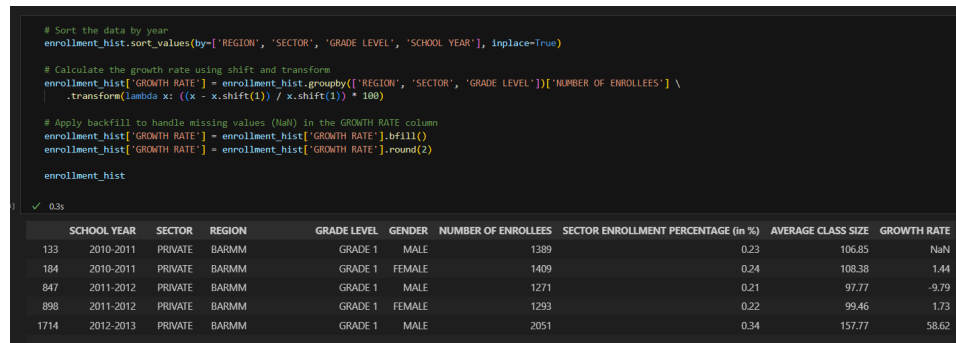


Figure 2.15: Growth Rate Column

C. Analysis Process and Techniques Used

C.1. Descriptive Analysis

For the descriptive analysis, the students' aim is to identify trends and outliers in educational enrollment trends across regions in the Philippines. Among the techniques used was *univariate analysis* in which key metrics such as enrollment, mean enrollment per region, and enrollment distributions by gender were computed, shown in previous sections. For example, the HistoricalEnrollmentData revealed enrollment disparities where specific regions exhibited significantly lower numbers of students enrolled, particularly in SUCs or LUCs. Additionally, *bivariate analysis* was also employed to show relationships between variables, such as gender and enrollment count, or sector and enrollment count, were explored using visualizations and correlation or heatmap analysis. Results showed gender balance in most regions but notable imbalances in the total number of enrollment in some rural regions. To visualize data for analysis, the authors' employed tools such as Python's Matplotlib and Seaborn libraries and Tableau to generate different diagrams and charts to represent trends.

Key findings in relation to enrollment trends show that public schools accounted for the largest count of enrollments, particularly in urban regions like NCR, which had over

64,000 Filipino students on average per school. Conversely, SUCs or LUCs had the lowest average enrollments, with a mean of 151 students per institution, possibly indicating accessibility issues. On gender, female and male enrollments were balanced overall. Findings in regional variations show that CAR exhibited the lowest enrollment figures, signaling a need for targeted interventions in the region. Also, some regions displayed zero enrollments in specific institutions, which indicate some data collection gaps in the original HistoricalEnrollmentData.

C.2. Exploratory Data Analysis and Visualization

Through exploratory data analysis (EDA) the authors aim to derive patterns, relationships, and trends within the HistoricalEnrollmentData dataset. This process includes creating the summary statistics, visualizations, and making some findings to provide meaningful insights in relation to educational access in the Philippines. Among the techniques and visualizations used are:

1. Regional Enrollment Analysis: Total number of enrollees by region was calculated using Tableau to provide geographical representation that can be interacted with.
2. Sector-Based Disparities: Heatmap of enrollment by gender and region was created to visualize gender-based trends across the country.
3. Grade-Level Insights: Stacked bar chart depicts the enrollment by grade level and gender, scaled to millions.
4. Temporal Trends: Line plots of enrollment trends across S.Y. 2010-2011 to S.Y. 2020-2021 for each region in the Philippines illustrated some variations in enrollment growth rates.

5. Growth Rates and Class Size: Bar charts showed average growth rate by region and depicted class size averages in the country.

C.3. Clustering

K-Means clustering is used in this analysis due to its ability to group data points based on similar characteristics without labeled data. This is helpful for finding enrollment patterns across *grade levels*, *regions*, or *sectors*. This allows the authors to identify regions with consistently low enrollment or disparities, providing valuable insights on the given data. The following steps are implemented:

C.3.a. Find optimal clusters using Elbow Method

Before applying the elbow method, the authors selected relevant features ('REGION,' 'SECTOR,' 'GRADE LEVEL,' and 'NUMBER OF EMPLOYEES') first. By doing this, the algorithm will be able to focus more on the relevant aspects of the data. Moving on to the elbow method, this was used to determine the optimal number of clusters by analyzing the within-cluster sum of squares (WCSS) for different cluster counts. Here, the authors found that the optimal number of clusters for the project is 3.

```
wcss = [] # Within-Cluster Sum of Squared Errors (WCSS)
for i in range(1, 11): # Test for 1 to 10 clusters
    kmeans = KMeans(n_clusters=i, random_state=42)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_) # Append WCSS value for each cluster count

# Plot the Elbow Method to find the optimal number of clusters
plt.figure(figsize=(8, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--', color='blue')
plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.title('Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')

# Identify and mark the "elbow" point
elbow = np.argmax(np.diff(np.diff(wcss))) + 2 # Optimal cluster point is where the second derivative is maximized
plt.plot(elbow, wcss[elbow - 1], 'o', color='orange') # Highlight the elbow point
plt.annotate(f'Optimal Clusters: {elbow}', xy=(elbow, wcss[elbow - 1]), xytext=(elbow + 1, wcss[elbow - 1] + 500),
            fontsize=9, bbox=dict(facecolor='lightgray'),
            arrowprops=dict(facecolor='black', arrowstyle='->'))

plt.show()
```

Figure 2.16 Elbow Method Implementation

C.3.b. Fit the KMeans Model and Assign Cluster Labels

In this step, the K-Means algorithm is applied using the optimal number of clusters that was obtained from the elbow method. Additionally, the visualization for the clusters are done here. This makes it easier to interpret the distribution of points and patterns present in the dataset.

```
optimal_clusters = elbow
kmeans = KMeans(n_clusters=optimal_clusters, random_state=42)
df['Cluster'] = kmeans.fit_predict(X_scaled)

fig = plt.figure(figsize=(10, 7))
ax = fig.add_subplot(111, projection='3d')

scatter = ax.scatter(
    df['REGION'], # REGION (Encoded numeric)
    df['NUMBER OF ENROLLEES'], # NUMBER OF ENROLLEES
    df['SECTOR'], # SECTOR (Encoded numeric)
    c=df['Cluster'], cmap='viridis', s=50, alpha=0.7
)

# Set titles and Labels
ax.set_title('K-Means Clustering (3D)')
ax.set_xlabel('Region')
ax.set_ylabel('Number of Enrollees')
ax.set_zlabel('Sector')

# Add colorbar to show cluster information
cbar = plt.colorbar(scatter, ax=ax, label='Cluster')
plt.show()

# Step 6: Inverse transform the encoded labels to human-readable labels for the summary
df['REGION_LABEL'] = le_region.inverse_transform(df['REGION'])
df['SECTOR_LABEL'] = le_sector.inverse_transform(df['SECTOR'])
```

Figure 2.17: Implementation of K-Means Clustering Algorithm

C.3.c. Cluster Summary

This section provides a detailed description on each cluster that was generated. Aside from the visualization, this also provides a good overview of the trends and areas that can be looked into about the data. Figure 2.18 shows a comprehensive cluster summary of this clustering.

```

# Initialize dictionaries to store data for all clusters
region_stats = {}
grade_stats = {}
sector_stats = {}
gender_stats = {}

df['REGION_LABEL'] = le_region.inverse_transform(df['REGION'])
df['SECTOR_LABEL'] = le_sector.inverse_transform(df['SECTOR'])

# Iterate through each cluster
for cluster in sorted(df['Cluster'].unique()):
    cluster_data = df[df['Cluster'] == cluster]

    # Top and bottom 5 regions
    top_regions = cluster_data['REGION_LABEL'].value_counts().head(5)
    bottom_regions = cluster_data['REGION_LABEL'].value_counts().tail(5)
    region_stats.append({
        "Cluster": cluster,
        "Regions with Most # of Enrollees": "<br>".join(top_regions.index),
        "Regions with Least # of Enrollees": "<br>".join(bottom_regions.index)
    })

    # Top and bottom 5 grade levels
    top_grades = cluster_data['GRADE_LEVEL'].value_counts().head(5)
    bottom_grades = cluster_data['GRADE_LEVEL'].value_counts().tail(5)
    grade_stats.append({
        "Cluster": cluster,
        "Grade Levels with Most # of Enrollees": "<br>".join("Grade " + str(grade) for grade in top_grades.index),
        "Grade Levels with Least # of Enrollees": "<br>".join("Grade " + str(grade) for grade in bottom_grades.index)
    })

    # Sector counts
    sector_counts = cluster_data['SECTOR_LABEL'].value_counts()
    sector_stats.append({
        "Cluster": cluster,
        "Public Sector": sector_counts.get(0, 0),

```

Figure 2.18 Cluster Summary Code Snippet

This summary highlights the *grade levels*, *genders*, *regions*, and *sectors* with different enrollment patterns, allowing for the identification of areas of concern for educational and resource allocation.

C.4. XGBoost

In order to provide additional insights and discover whether the *sectors*, *grade levels*, *regions*, and *gender* will be a factor in the number of enrollees, the XGBoost model is utilized. XGBoost, which stands for Extreme Gradient Boosting, is a highly effective machine learning tool known for its ability to process large datasets quickly and accurately. It works by combining multiple decision trees and using a technique called gradient boosting to refine its predictions (Chen & Guestrin, 2016). With XGBoost, the patterns in the data, and key factors influencing enrollment will be uncovered and used to make reliable predictions.

C.4.a. Model Training and Testing

The XGBoost model undergoes training and testing with a total of 200 epochs. It uses the 'SECTOR,' 'GRADE LEVEL,' 'GENDER,' and 'REGION' as inputs in order to predict the number of enrollees. For the metrics, the root mean square error or RMSE is employed for both training and testing in order to track how the model performs.

To verify if the model is performing well, the students visualized the actual vs predicted number of enrollees and the values appeared to be close, indicating the model's ability to predict well.

C.4.a Model Application

After training and testing the XGBoost model, we used it to predict the number of enrollees for each cluster. The inputs included the top five *regions*, *sectors*, *grade levels*, and *genders* for each cluster to ensure the predictions were accurate. This application of the model is crucial for assessing whether the enrollment numbers for each cluster will remain consistent in the coming years, from 2025 to 2029. Understanding these trends could offer valuable insights into which areas the government should observe and prioritize for educational planning.

Chapter III Results

A. Visualization and Their Interpretations

A.1. Total Enrollment by Region (2010-2011 to 2020-2021, in Millions)

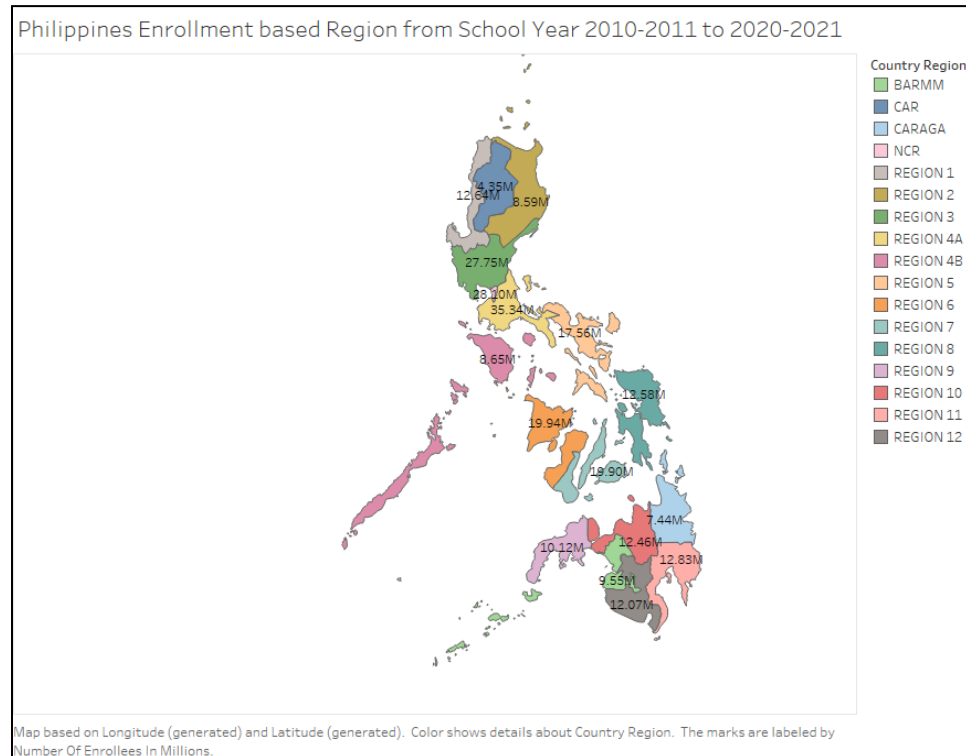


Figure 3.1: Tableau Map Distribution Visualization

The map visualization of the Philippines with the total number of enrollees by region from School Year 2010-2011 to 2020-2021, shows a significant difference in enrollment numbers across different regions, with CAR having low enrollment rates and CARAGA following as the second lowest with 7.44 million enrollees. In contrast, NCR and Region 4A have the highest enrollment numbers, with NCR standing out due to its densely packed number of enrollees despite being the smallest region in terms of land area. This imbalance highlights underlying issues such as population density,

accessibility, and resource allocation, which may affect enrollment distribution. It is recommended for the government to focus on balancing the resource allocations and improving educational infrastructure to all regions, hoping to further improve the balance of the number of enrollees all throughout the regions in the Philippines.

A.2. Sector-Wise Enrollment by Region (2010-2011 to 2020-2021)

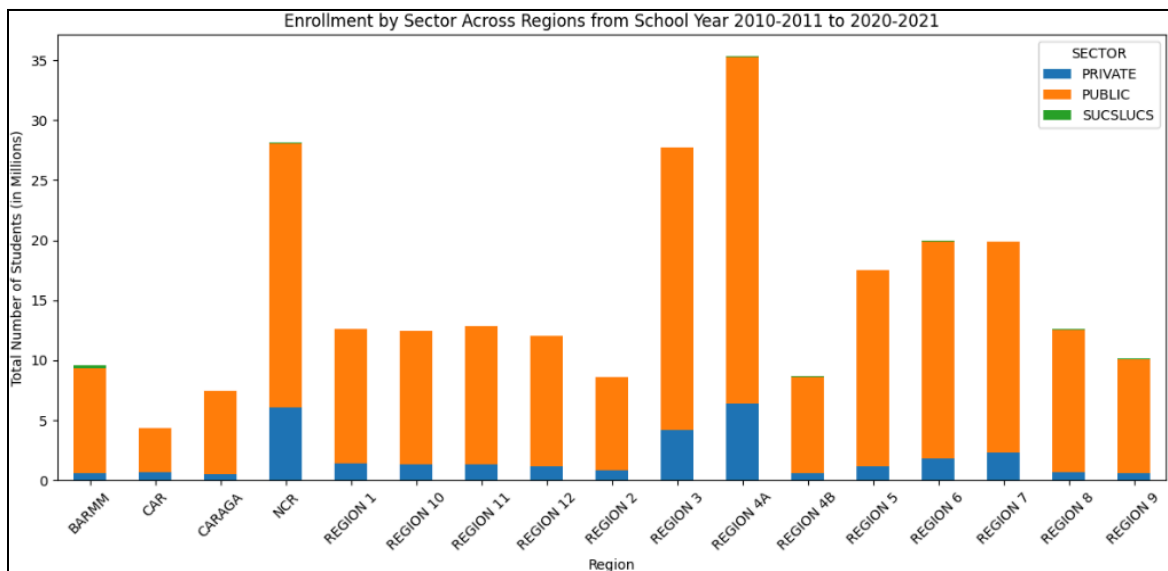


Figure 3.2: Enrollment by Sector Across Regions

Figure 3.2 Shows the bar plot of enrollment by sector across regions starting from the academic year 2010-2011 up until 2020-2021. It shows that there are a lot of students that study under the Public sector overpowering the Private and SUCs or LUCs sector. Given that Region 3 has the highest enrollment percentage following Region 4A these high figures in regions suggest a strong reliance on public education due to higher population density or affordability factors. Moving forward to some urban regions like NCR and Region 4A likely see a higher private school enrollment due to a concentration of middle- to high-income families and a variety of private education options. In BARMM there is a lower private school availability due to economic and geographic

constraints. Also, Heavy reliance on government-supported education systems in underserved regions. Urbanized regions such as NCR, Region 4A, and Region 3 show higher private school enrollments, reflecting economic advantages and access to diverse educational opportunities. It displayed that in regions CAR, BARMM, and Caraga they are more rural and less developed regions resulting in lower private school enrollments and greater dependence on public schools.

A.3. Sectoral Distribution

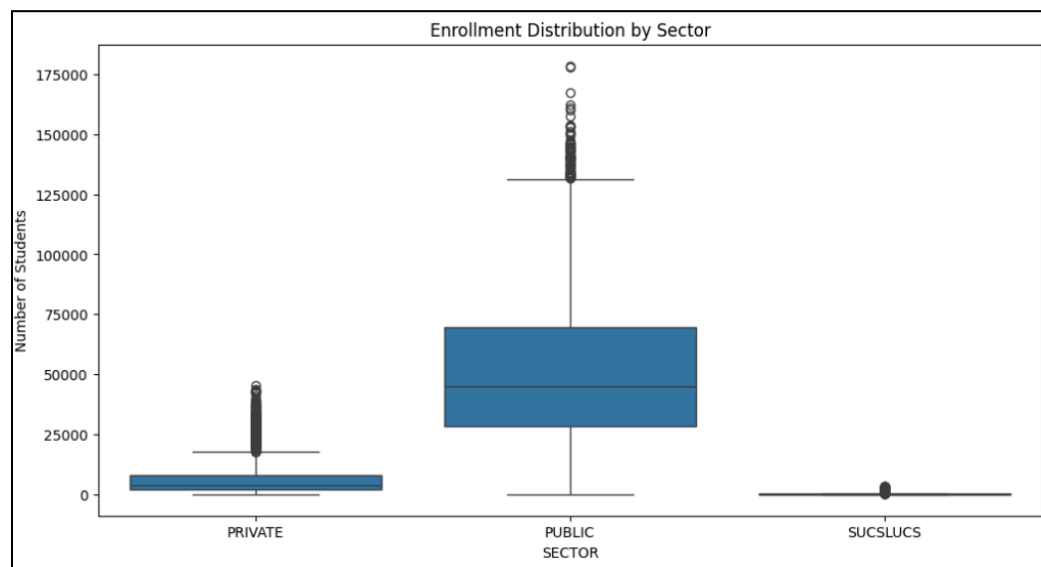


Figure 3.3: Enrollment Distribution by Sector

The visualization in Figure 3.3 suggests that the majority of enrollees come from the public sector, reflecting the economic reality that many Filipinos cannot afford private school tuition. This highlights the crucial role public schools have in providing accessible education to the broader population. Given this context, it is highly recommended that the government prioritize the improvement of public school infrastructures to accommodate the growing demand and ensure a quality education for all. Strengthening public

education systems will ensure a more inclusive and equitable educational landscape for all people given the current poverty status in the Philippines.

A.4. School Year Enrollee Trend based on Regions (in Millions)

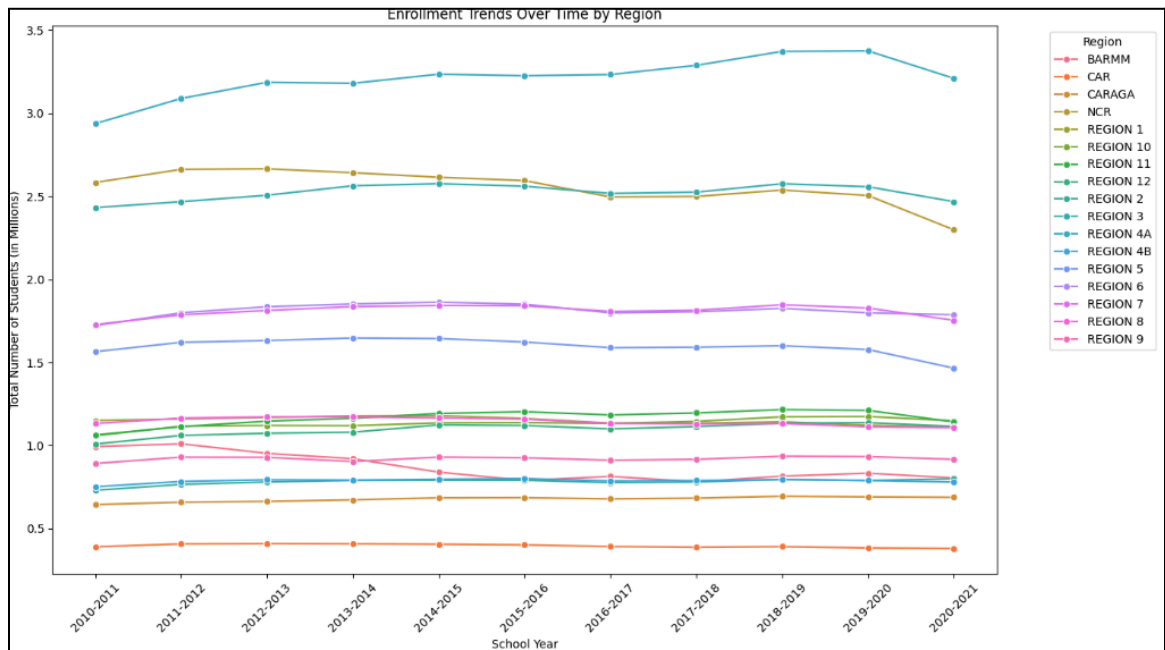


Figure 3.4: Enrollment Trends over School Year by Region

The chart reveals that Region 4A consistently has the highest enrollment numbers over the school years, likely due to its high population density and rapid urbanization, which contribute to better access to education. NCR follows closely, maintaining significant enrollment levels due to its urban setting and abundant educational facilities. Moderate increases can be seen in regions such as Region 7 and Region 6, while smaller regions like CAR and BARMM remain at the lower end, indicating challenges in access to educational resources. The slight declines observed from 2019-2020 to 2020-2021 across most regions might be attributed to external disruptions like the COVID-19 pandemic. It is recommended that it is better to strengthen education access to smaller regions through investing educational resources and facilities to address enrollment gaps.

Moreover, In regions such as NCR or Region 4A with a high number of enrollees, there should be funding for additional infrastructure to prevent overcrowding and maintain the quality education. And as mentioned COVID-19 could affect the number which there should be flexible modalities to address the enrollment declines.

A.5. Enrollment by Gender and Region (in Millions)

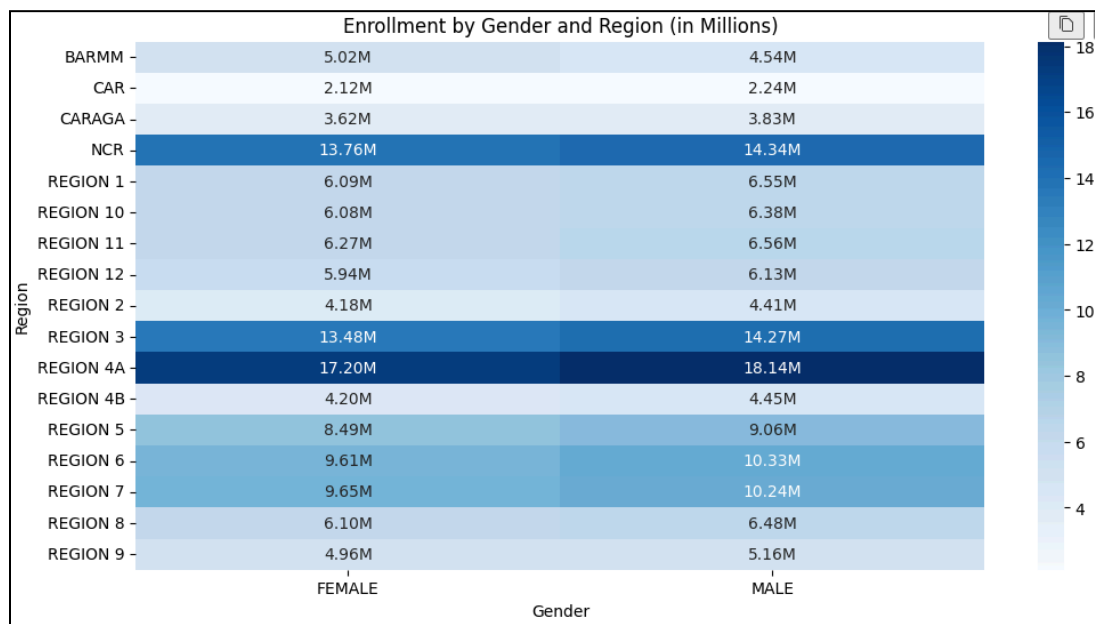


Figure 3.5: Enrollment by Gender and Region (in Millions)

The heatmap visualization of total enrollees from the school years 2010-2011 to 2020-2021, categorized by gender and region in Figure 3.5, reveals that there are slightly more male enrollees than females across most regions. However, it still shows a relatively balanced gender distribution, which is a positive indicator of gender equality in the Philippines. Among the regions, BARMM is the only region with a higher number of females than male. Even with only a slight imbalance on the distribution of genders, it is still recommended to promote and implement gender equality programs to ensure that

both genders have equal access to education opportunities, this will help further strengthen efforts toward achieving gender equality in education across all regions.

A.6. Total Enrollees by Grade level and Gender (in Millions)

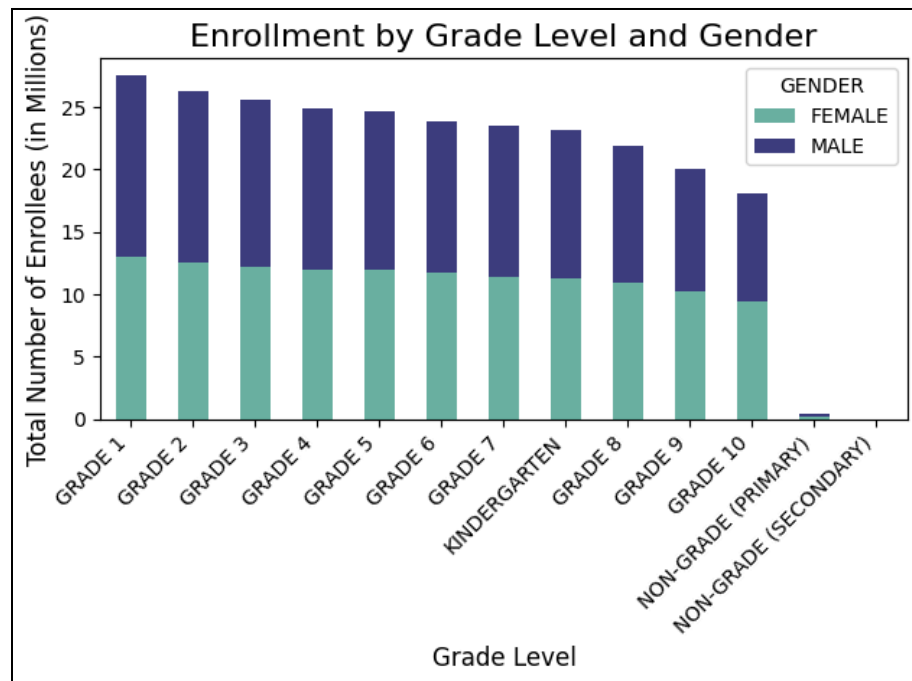


Figure 3.6: Enrollment by Grade Level and Gender

The data reveals a significant gender disparity in enrollment, with female students dominating the early grades but declining as they progress. Additionally, enrollment numbers decrease significantly as students move from elementary to secondary levels, particularly in Grades 9 and 10. This suggests potential retention challenges at the secondary level. Non-Grade primary and secondary level has the least number of enrollees, overall. To address these issues, targeted outreach and support programs for female students, enhanced support services for all students, and optimization of non-grade programs are recommended. By implementing these strategies, educational institutions can improve student retention, promote gender equity, and create a more inclusive learning environment.

A.7. Average Class Size by Region

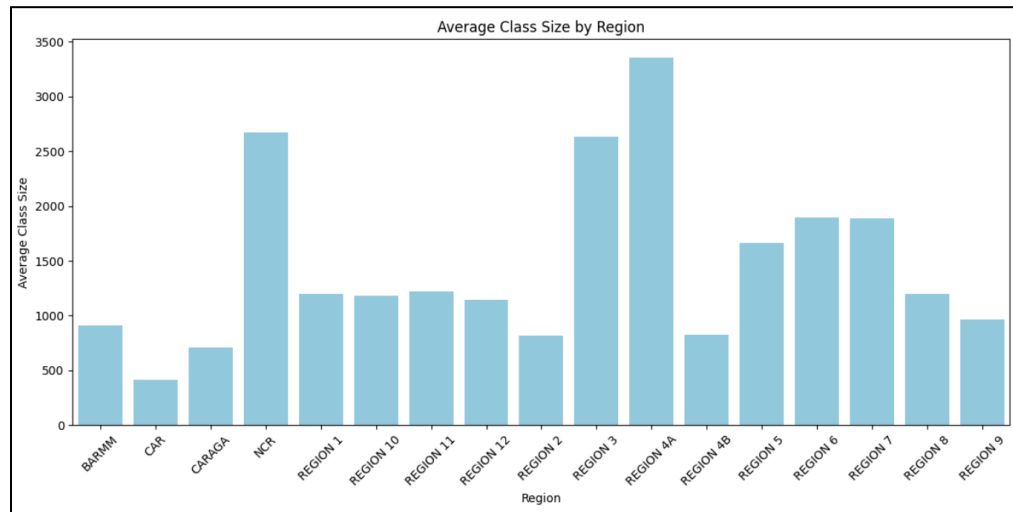


Figure 3.7: Average Class Size by Region

The figure above illustrates the average class size across regions, calculated by dividing the number of enrolled students by the number of grade levels. Region 4A and NCR have the highest average class sizes, while CAR and CARAGA have the lowest. This highlights the unequal distribution of educational resources and infrastructure, often concentrated in urbanized regions like NCR and Region 4A. To promote balanced growth, the government needs to prioritize investments in educational facilities and workforce across all regions, ensuring that all areas receive adequate support. By addressing this, educational development can be distributed more evenly focusing on nationwide progress.

A.8. Sector Enrollment Percentage by Region

Figure 3.8, highlights sector enrollment trends over the school years. The private sector shows a relatively stable percentage, with a decline observed from 2019 to 2021. The public sector experienced a notable decline between 2013 and 2015, while the SUCs

or LUCs sector showed a gradual decline from the 2012-2013 academic year to 2015-2016. Although SUCs or LUCs partially recovered afterward, it remains below the private and public sectors.

A significant drop across all sectors around 2014-2015 suggests potential factors like policy changes, funding issues, or external influences affecting enrollment. Post-2016-2017, enrollment percentages appear to stabilize, possibly due to improved education plans and opportunities encouraging student enrollment. To address challenges under the SUCs or LUCs sector it is better to increase the allocation of funds and provide support that could help boost enrollment. While for the public sector, an enhanced quality of education and expanded scholarship program could attract more students.

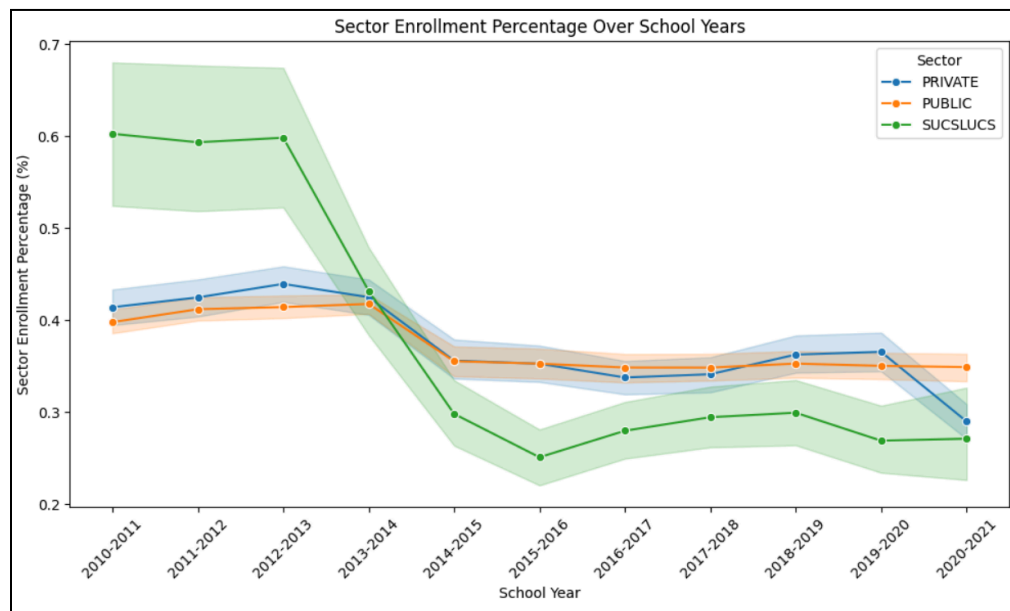


Figure 3.8: Sector Enrollment Percentage

B. Machine Learning Model Implementation and Evaluation

B.1. K-Means Model

The K-Means clustering algorithm is an unsupervised machine learning technique that partitions data into k clusters based on feature similarity. It is simple and computationally efficient, which makes K-Means a commonly used method for exploratory data analysis in large datasets (Jain, 2010). For the analysis of historical enrollment data in the Philippines, K-Means clustering provides a systematic way to identify patterns across regions, sectors, and genders.

B.1.a. Elbow Method

For the K-Means clustering algorithm, it would require the selection of the number of optimal clusters, denoted as k . To determine its optimal value, the *elbow method* was employed. This method evaluates the within-cluster sum of squared errors (WCSS) across a range of potential counts of clusters. The aim is to identify a point where the reduction in WCSS begins to diminish, indicating the most appropriate number of clusters.

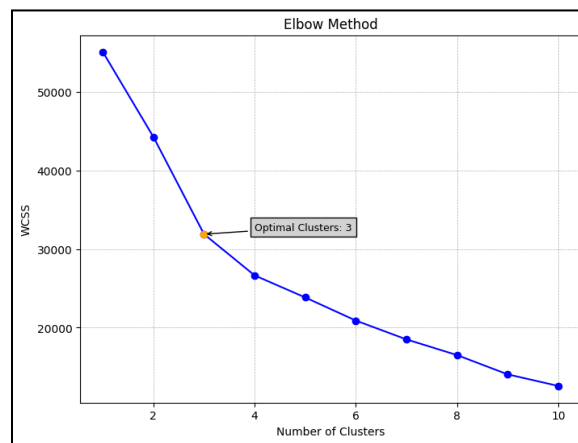


Figure 3.9: Elbow Method Result

A range of k values from 1 to 10 were tested using the dataset, scaled for consistent measurement across features. For each k , the K-Means algorithm was trained, and the WCSS was calculated. These WCSS values were plotted against the corresponding cluster counts. The plot generated by the elbow method is shown in Figure 1. The graph exhibits a sharp decline in WCSS as k increases, but the rate of decrease slows notably after $k = 3$. This *elbow* point, marked on the plot, represents the optimal count of clusters.

B.1.b. K-Means Clustering

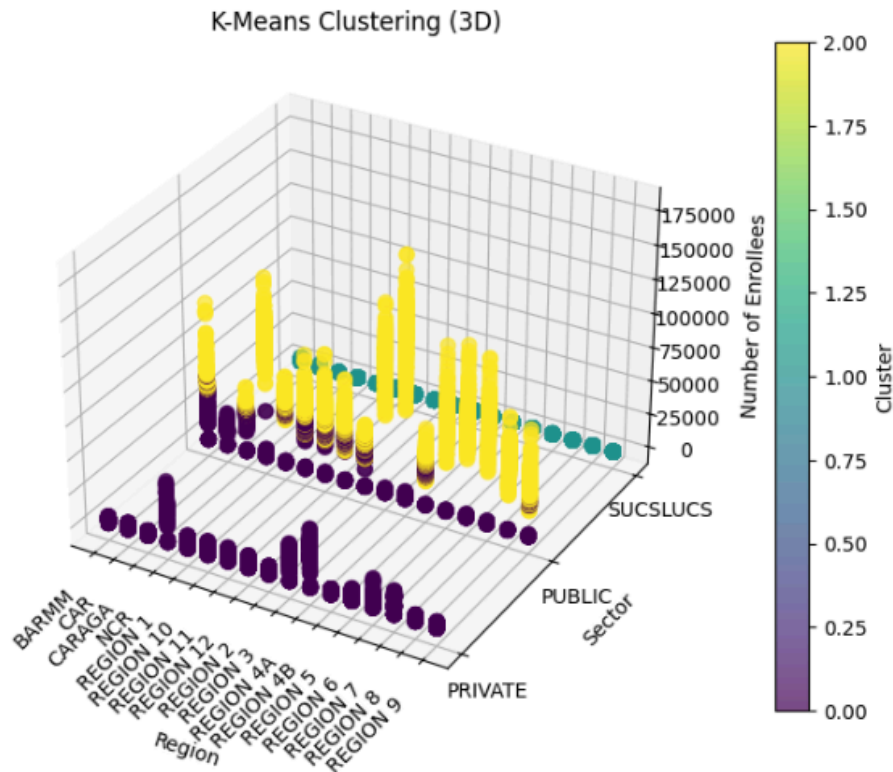


Figure 3.10: 3D K-Means Clustering

This visualization of K-Means clustering shows a distinct pattern in enrollment distribution across the three clusters. The enrollees from *Cluster 0* come from both private and public sectors. Their average number in comparison to *Cluster 2*, which has

the largest number of enrollees, may signify that its lower number may be due to the exclusivity of private and specialized public schools in comparison to normal public ones. Additionally, the presence of tuition fees in private institutions may also affect its accessibility to the masses. The mix of private and public sectors here also suggests that it has a more diverse demographic.

As for *Cluster 1*, it has the lowest number of enrollees, all of which are from SUCs or LUCs. The severely low number may be due to the limited capacity of these institutions as well as their strict enrollment criteria. Additionally, geographic accessibility due to fewer available campuses may also contribute to these numbers. This highlights the need for more state universities and colleges to meet the demand.

Lastly, *Cluster 2* has the highest enrollment among the three. This cluster is entirely made up of public schools. The high numbers may be due to the wide availability as well as affordability of public education. Similarly, overcrowding due to the dense population may also be a factor.

B.1.c. K-Means Clustering with Principal Component Analysis (PCA)

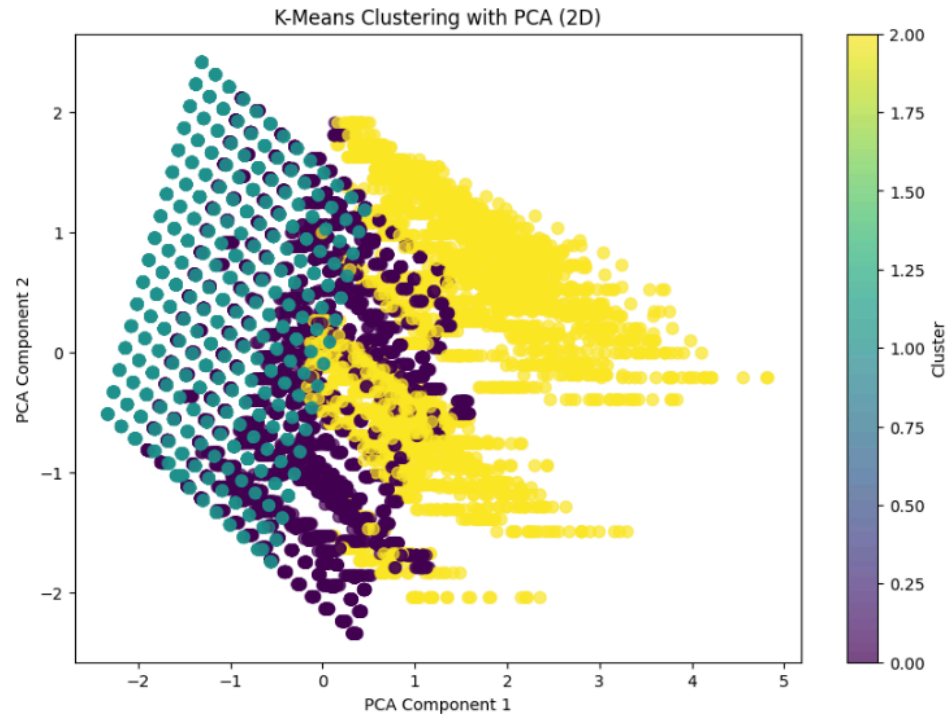


Figure 3.12: 2D K-Means Clustering with PCA

Figure 3.12 shows a clear separation of the clusters along each PCA component. Before further exploration of each cluster, the figure below shows how each principal component (PC) is influenced by the features. This will provide better insight into how each component describes the clusters.

```
# Explained variance ratio for the PCA components
explained_variance = pca.explained_variance_ratio_
print("\nExplained Variance Ratio:")
print(f"PC1: {explained_variance[0]:.2f}, PC2: {explained_variance[1]:.2f}")

# PCA Loadings (Feature Contributions)
loadings = pca.components_
print("\nPCA Loadings:")
for i, component in enumerate(loadings, start=1):
    print(f"PC{i}: {dict(zip(features, component))}\n")

Explained Variance Ratio:
PC1: 0.30, PC2: 0.25

PCA Loadings:
PC1: {'REGION': 0.3125126886566356, 'SECTOR': -0.35087175458848163, 'GRADE LEVEL': -0.5284173362623765, 'NUMBER OF ENROLLEES': 0.7071067811865472}
PC2: {'REGION': -0.241883569886627, 'SECTOR': 0.7362919553032281, 'GRADE LEVEL': -0.6319545040385819, 'NUMBER OF ENROLLEES': -2.6645352591003757e-15}
```

Figure 3.13: Variance Ratio

The explained variance in PCA refers to the total variability in the data that each PC captures. This tells how much information is retained. In this clustering, the total

variance captured was 55% of the total variance in the dataset. With 30% coming from PC1, and the remaining coming from PC2. According to the loadings, PC1 is highly influenced by ‘NUMBER OF ENROLLEES’ having a value of 0.71 while PC2 is mostly focused on ‘SECTOR’ getting a value 0.73.

Going back to Figure 3.10. The points in *Cluster 0* being located in the upper left quadrant imply schools or regions with average enrollment numbers due to having a more targeted sector type. Despite this, it’s not enough to have as low enrollment as the other cluster. As for *Cluster 1*, it has $PC1 < 0$ as well as $PC2 < 0$. This reflects a very low enrollment number as well as an extremely niche sector such as SUCs or LUCs. Lastly, *Cluster 2* spans both quadrants 1 and 4. This suggests an overlap in the number of enrollees, but a difference in sector type. Overall, this cluster has a high enrollment number but has a diverse set of institutions where some are more traditional and some are more specialized.

B.1.d. Cluster Summary

The K-Means clustering analysis grouped the regions into three distinct clusters based on enrollment patterns, sector types, and other demographic dimensions. Table 3.1 shows the summary.

Table 3.1: K-Means Cluster Summary

	Cluster		
	0	1	2
Region Count	6,029	4,590	3,151
Average Enrollment	9,344.37	151.06	64,372.75
Total Enrollment	56,337,219	693,365	202,838,547

Min. Enrollment	0	0	17,983
Max. Enrollment	49,042	3,358	178,471
Regions	<i>All Regions</i>	<i>All Regions</i>	<i>All Regions</i>
Sectors	Public & Private	SUCS LUCS	Public
Genders	Male & Female	Male & Female	Male & Female

For *Cluster 0*, it encompasses both private and public schools, with enrollment figures spanning all regions in the dataset. It has a moderate average enrollment of approximately 9,344 students per institution. Notable patterns show that schools with a wide range of enrollment values, from zero to a maximum of 49,042 students. These regions likely represent areas with moderately developed education systems, balancing accessibility and enrollment across public and private sectors.

For *Cluster 1*, it consists predominantly of State Universities and Colleges (SUCs) and Local Universities and Colleges (LUCs), with very low average enrollment figures (151 students per institution). Some notable patterns show the smallest total and average enrollment values across clusters, with a maximum enrollment of only 3,358. Includes all regions but represents institutions or sub-sectors facing accessibility issues, possibly due to limited infrastructure, resources, or low demand even. This could imply that such areas require focused intervention, such as capacity-building programs, resource allocation, or infrastructure development to improve enrollment rates.

Lastly, *Cluster 2* represents public schools with the highest average and total enrollment figures. Institutions in this cluster cater to large student populations. Average enrollment is significantly high at 64,373 students, with a maximum of 178,471 per institution. It likely reflects urbanized regions with high population density, benefiting

from established public education infrastructure. However, these regions may face issues like overcrowding and strained resources, necessitating interventions for quality improvement rather than accessibility.

In addition to the comprehensive summary provided in Table 3.2, the students also conducted a detailed cluster analysis, focusing on identifying the regions and grade levels with the highest and lowest enrollment numbers per cluster, as well as the sector and gender distributions within each cluster. These analyses contribute to a deeper understanding of the clusters.

Table 3.2: Comparison of Regions per Cluster

Cluster	Regions with Most # of Enrollees	Regions with Least # of Enrollees
0	CAR	Region 4A
	CARAGA	Region 5
	BARMM	Region 6
	Region 2	Region 7
	Region 4B	Region 8
1	BARMM	Region 10
	Region 3	Region 1
	Region 8	NCR
	Region 7	CARAGA
	Region 6	Region 9
2	Region 3	Region 10
	Region 4A	Region 4B
	Region 5	Region 2
	Region 6	BARMM
	Region 7	CARAGA

Table 3.2 compares the regions with the highest and lowest number of enrollees across three clusters. In Cluster 0, the region with the most enrollees is CAR, while

CARAGA and BARMM have the fewest enrollees. Other regions with low enrollment in this cluster include Region 2, Region 4B, Region 4A, Region 5, Region 6, Region 7, and Region 8. For Cluster 1, BARMM has the highest enrollment, and the regions with the least number of enrollees are Region 3, Region 8, and Region 7, along with Region 6, Region 10, Region 1, NCR, and CARAGA. In Cluster 2, the regions with the most enrollees are Region 3, Region 4A, and Region 5, while BARMM and CARAGA have the fewest. Additional regions with low enrollment in this cluster include Region 4B, Region 2, Region 6, Region 7, and Region 10.

Table 3.3: Comparison of Grade Levels per Cluster

Cluster	Regions with Most # of Enrollees	Regions with Least # of Enrollees
0	Grade 9	Grade 5
	Grade 8	Grade 4
	Grade 10	Grade 3
	Grade 7	Grade 2
	Non Grade (Primary)	Kindergarten
1	Kindergarten	Grade 8
	Grade 1	Grade 9
	Grade 2	Grade 10
	Grade 3	Non Grade (Primary)
	Grade 4	Non Grade (Secondary)
2	Kindergarten	Grade 1
	Grade 2	Grade 7
	Grade 3	Grade 10
	Grade 4	Grade 8
	Grade 5	Grade 9

Based Table 3.3, the grade levels with the most enrollees in Cluster 0 are Grade 9, followed by Grade 8 and Grade 10, while the grades with the least enrollees are

Kindergarten, Grade 2, Grade 3, Grade 4, Grade 5, and Non-Grade (Primary). For Cluster 1, Kindergarten has the highest enrollment, with Grade 1, Grade 2, Grade 3, Grade 4, Grade 8, Grade 9, and Grade 10 following closely behind, while Non-Grade (Primary) and Non-Grade (Secondary) have the least enrollees. In Cluster 2, Kindergarten has the highest enrollment, followed by Grade 2, Grade 3, Grade 4, and Grade 5, while Grade 1, Grade 7, Grade 10, Grade 8, and Grade 9 have the least enrollees.

Table 3.4: Comparison of Number of Enrollees in each Sector per Cluster

Cluster	Public	Private	SUCs or LUCs
0	1439	4590	0
1	0	0	4590
2	3151	0	0

Table 3.4 shows a comparison of the number of enrollees in each sector per cluster. Cluster 0 has enrollees split between private and public sectors. This shows a more diverse education option, with a preference for the private sector. Whereas Cluster 1 is fully concentrated in the SUCs or LUCs sector. This suggests that this cluster is a representative of regions or groups with reliance on state-funded colleges and universities. Lastly, Cluster 2 has a significant concentration of enrollees in the public sector. This represents regions that are more into public education, possibly due to cost or widespread availability.

Table 3.5 shows the gender distribution for every cluster. Overall, all clusters have almost equal distributions which may suggest that gender is not a significant factor in enrollment disparities across these clusters.

Table 3.5: Comparison of Gender Distribution per Cluster

Cluster	Male	Female
0	3019	3010
1	2295	2295
2	1571	1580

B.1.e. Comparison of K-Means Clustering

The analysis includes two variations of K-Means clustering that were employed to classify the dataset into clusters based on enrollment patterns: (1) 3-dimensional clustering using the original features and (2) 2-dimensional clustering after dimensionality reduction using PCA. The comparison between these two approaches highlights their respective advantages and limitations.

Table 3.6: Strengths and Limitations: 3D K-Means Clustering

3-Dimensional K-Means Clustering	
Strengths	Weaknesses
Maintains the original feature dimensions, ensuring that clustering is based on the raw relationships within the data.	The visualization is less intuitive due to the 3D nature, which can be harder to interpret in static graphs.
Suitable for datasets with a manageable number of features.	As the number of features grows, visual representation becomes impractical.

First, the 3-dimensional K-Means clustering, illustrated in Figure 3.6, uses three primary features—region (encoded), number of enrollees, and sector (encoded)—as dimensions for clustering. To visualize, a 3-dimensional scatter plot is made, wherein the three axes correspond to the encoded features mentioned earlier. Clusters were

represented using distinct colors. Table 3.6 highlights the clustering’s strengths and limitations.

Second, the 2-dimensional K-Means clustering, illustrated in Figure 3.12, employed Principal Component Analysis (PCA) was applied to reduce the dataset’s dimensions to two principal components. Clustering was then performed on the PCA-reduced dataset. To illustrate the clusters, a 2D scatter plot was used to display the clusters, with each axis representing a principal component. Similarly, clusters were represented with their corresponding colors. Table 3.7 provides the clustering’s strengths and limitations.

Table 3.7: Strengths and Limitations: 2D K-Means Clustering

2-Dimensional K-Means Clustering	
Strengths	Weaknesses
Simplifies the dataset by capturing most of the variance in just two dimensions, improving visualization and interpretability.	Dimensionality reduction leads to loss of information, as PCA does not retain all the original feature relationships.
Ideal for communicating results to non-technical audiences.	Interpretation of the principal components is less intuitive compared to the original features.

In relation to clustering quality, both methods identified similar clusters, indicating consistency in the patterns of enrollment disparity. However, the 3D clustering retained more nuanced relationships, while the 2D clustering provided a generalized view. For interpretability, the PCA-based model was easier to interpret and visualize but lacked the depth or richness of the 3D clustering to represent individual feature contributions. Whereas for its applicability, the 3D model is more suitable for in-depth

analyses where maintaining original feature relationships is critical, especially in the context of enrollment data. In contrast, the PCA-based approach suits for presentation purposes.

B.2. XGBoost Model

A supervised model, XGBoost, is also utilized to gain an insight into whether these clusters will continue to have the same pattern in the future in terms of the number of enrollees. The XGBoost model uses ‘SECTOR,’ ‘GRADE LEVEL,’ ‘REGION,’ and ‘GENDER’ as inputs and the ‘NUMBER OF ENROLLEES’ as its output. It was set to run for 500 boost rounds with an early stopping of 30 if the model shows no progress over rounds.

B.2.a. Evaluation

Table 3.8: XGBoost Model’s Optimal Parameters

Parameters	
max-depth	7
eta (learning rate)	0.1
subsample	0.8
colsample_bytree	1.0
min_child_weight	3
lambda (L2 regularization)	2
alpha (L1 regularization)	2
num_boost_round (after early stopping)	99

The XGBoost model was configured with several hyperparameters to optimize its performance. The max depth of the decision trees was set to 7, which helps control the complexity and prevents overfitting by limiting the tree's growth. The eta, or learning rate, was set to 0.1, allowing for a more gradual and precise optimization process during training. The subsample value was set to 0.8, meaning 80% of the training data was used for each tree, which helps reduce overfitting by adding randomness. The colsample_bytree parameter was set to 1.0, indicating that all features were used in constructing each tree. To ensure that the trees do not overfit the training data, the min_child_weight was set to 3, imposing a threshold for the minimum sum of instance weights in each child. The model also employed L2 regularization (lambda) and L1 regularization (alpha), both set to 2, to penalize large model coefficients and encourage sparsity. Finally, the num_boost_round ended at 99, meaning the model was trained for 99 boosting rounds given that early stopping was applied to prevent overfitting. These hyperparameters collectively help ensure that the model remains efficient, balanced, and generalizable to unseen data.

Table 3.9: Preview of Model's Performance across Boost Rounds

XGBoost Model Performance		
Epoch/Round	Training RMSE	Testing RMSE
0	26,529.81463	26,210.54236
9	11,613.57052	11,507.13216
39	4,071.00374	4,160.93928
69	3,863.44561	4,037.09272
99	3,835.20321	4,060.69254

The table shows the training and testing Root Mean Squared Error (RMSE) for each epoch/round during the training of the model. RMSE is a suitable metric for this case given that it provides an estimate of the average discrepancy between the predicted and actual values, reflecting how far off, on average, the model's predictions are from the true outcomes in the dataset (Deepchecks, 2024).

Initially, at round 0, the training RMSE is 26,529.81, and the testing RMSE is 26,210.54, indicating significant error during the early stages of training. As the training progresses, the RMSE values decrease, which is expected as the model learns to make better predictions. By round 9, the RMSE for training drops to 11,613.57, and the testing RMSE decreases to 11,507.13, showing improvement. By round 39, the training RMSE has decreased further to 4,071.00, while the testing RMSE is 4,160.94, which shows the model is nearing a state of better generalization. The best round is at round 69, where the training RMSE improves to 3,863.45, while the testing RMSE is 4,037.09. Finally, at round 99, the training RMSE reaches 3,835.20, and the testing RMSE slightly increases to 4,060.69, which urges the model to stop early.

The decrease in RMSE during the earlier rounds signifies the model's improvement in accuracy, while the stabilization or slight increase in RMSE in later rounds can highlight the trade-off between underfitting and overfitting. Overall, RMSE around 3,800 to 4,000 is generally good and is considered a small relative error given that the maximum number of enrollees originally ranges around 180,000.

B.2.b. Prediction vs Actual

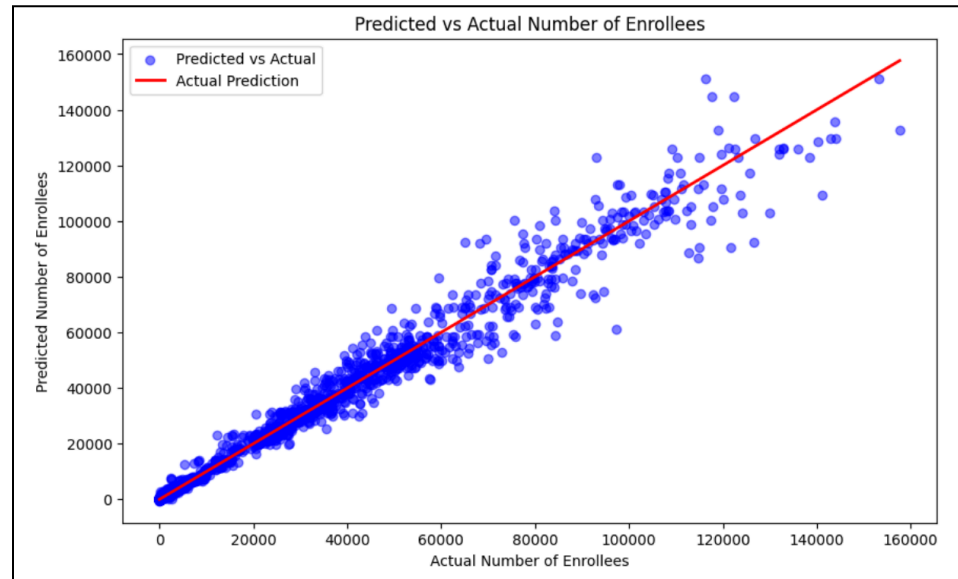


Figure 3.14: Predicted vs Actual Number of Enrollees using the XGBoost Model

Figure 3.14 illustrates a comparison between the actual enrollment numbers and those predicted by the XGBoost model. As shown, the predicted values closely align with the actual values, suggesting that the model's performance is quite accurate. This close correlation implies that the XGBoost model, which has been thoroughly trained and tested, can be considered reliable for making future predictions.

B.2.c. Predicted Number of Enrollees in Future School Years

In section A of this chapter, the use of K-means clustering is demonstrated. The results of that model revealed that there are 3 clusters with patterns and features closely related. Given that the data is only until 2021, the XGBoost model is used to predict the number of enrollees in each cluster data on the upcoming five school years: 2022-2023, 2023-2024, 2025-2026, and 2026-2027. In order to avoid bias in predictions, the values utilized for 'REGION,' 'SECTOR,' 'GENDER,' and 'GRADE LEVEL' were ones with

highest values in each cluster. Using these values will give information on what will be the number of enrollees in each cluster given the same *regions*, *sectors*, and *grade levels*.

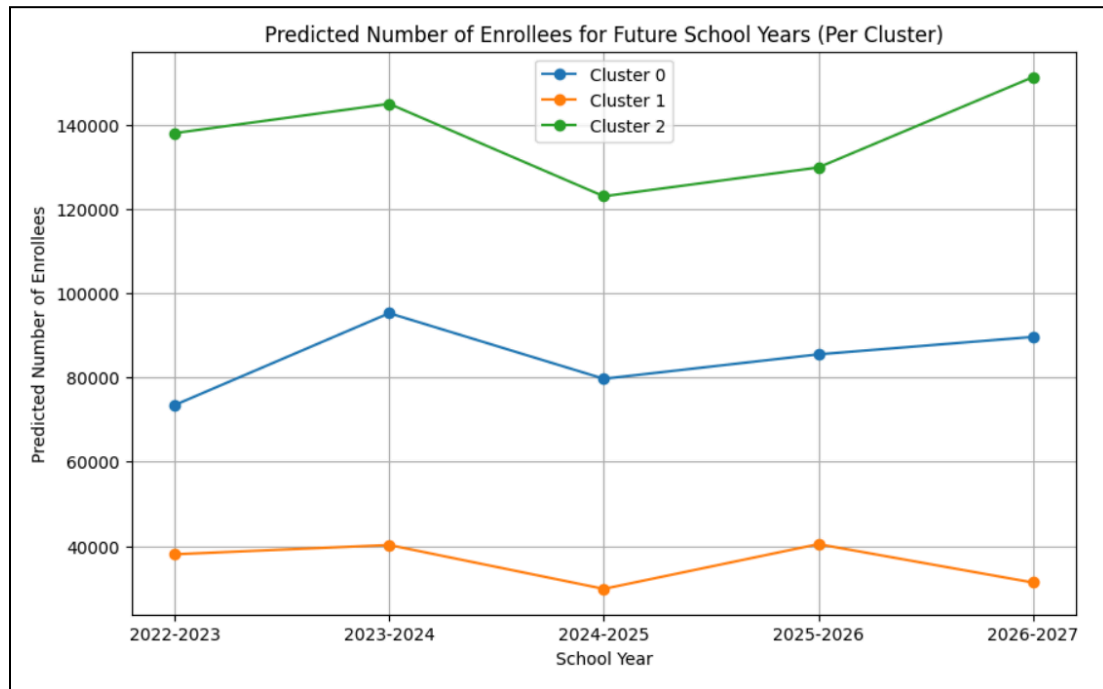


Figure 3.15: Predicted Number of Enrollees for Future School Years (Per Cluster)

As shown in Figure 3.15, Cluster 2 will have the highest number of enrollees, followed by Cluster 0, and the least is Cluster 1. This prediction is the same in the analysis of clusters predicted in section A, implying that given the same circumstances, Cluster 2 or the Public Schools in Region 3, Region 4a, and Region 5 will continue to increase over years. While Cluster 1, or the SUCS LUCS Schools mostly on Region 4A, NCR, and Region 3 will continue to be much lower given the observed pattern in the data seen by the XGBoost model. Given these predictions, it might be necessary to focus more on areas and observe the factors contributing to low enrollment count, and also investigate if there are negative outcomes brought by the very high enrollment count in public schools in Cluster 2.

Chapter IV Recommendations and Conclusions

A. Summary

The analysis primarily focuses on the `HistoricalEnrollmentData` dataset, a record of enrollment trends in Philippine elementary and junior high schools spanning S.Y. 2010-2011 to S.Y. 2020-2021. Sourced from the Department of Education (DepEd) and Open Data Philippines, the dataset includes key dimensions such as *regions*, *sectors*, *grade levels*, and *gender*, providing a foundation for understanding disparities in educational access. The analysis of the authors addresses and gives some insights of factors like poverty, inadequate infrastructure, and the digital divide, which continue to limit educational opportunities for Filipino learners. Through detailed descriptive analysis, the study identifies patterns and outliers, setting the stage for targeted interventions to address these disparities.

Also, a structured data pipeline was implemented to ensure data integrity and reliability before analysis. This process involved inspecting, cleaning, and validating datasets from both educational levels, addressing inconsistencies, and resolving anomalies. Outliers were identified using the Interquartile Range (IQR) method, but were retained to preserve critical variations in enrollment trends from the dataset. The authors kept these outliers to ensure that the analysis captures unique enrollment patterns, particularly in regions possibly facing extreme challenges. After reclassifying categorical variables and merging the datasets, the final consolidated dataset facilitated exploratory analysis and advanced techniques like K-Means clustering and XGBoost predictions. Overall, the authors employed these methods and insights into regional, sectoral, and demographic disparities, with the goal to lay groundwork for actionable recommendations to improve educational equity in the Philippines.

B. Conclusion

Based on the analysis conducted and the visualizations created to understand the data on the Enrollment of Elementary and Junior High School students, the following insights were attained:

B.1. Regions

The data reveals that NCR and Region 4A consistently have the highest enrollment numbers. It is concluded that NCR despite being relatively small in land area resulted in a large number of enrollees. This highlights underlying challenges such as high population density, uneven resource allocation, and accessibility issues, which may strain education infrastructure and affect the equitable distribution of enrollees. On the other hand, CAR and CARAGA exhibit the lowest enrollment figures, reflecting gaps in educational access and infrastructure in these less urbanized regions. Furthermore, the slight decline in enrollment was observed from 2019-2020 to 2020-2021 across most regions which likely be attributed to disruptions from the COVID-19 pandemic, affecting students' ability to access education. With that being said, the government should invest in growing regions to expand schools and allocate resources to manage the high demand and prevent overcrowding. While those underserved areas, improve the accessibility in education through building schools, enhancing transportation systems, and introducing digital learning in the remote areas.

B.2. Sectors

The data reveals that NCR and Region 4A experience higher private school enrollments, likely due to the concentration of middle- to high-income families and the availability of various private education options. Region 4A also shows strong public

sector enrollment, possibly driven by high population density and the affordability of public education. Urbanized regions like NCR, Region 4A, and Region 3 have a higher share of private school enrollees, reflecting economic advantages and greater access to diverse educational opportunities. In contrast, rural and less developed regions like CAR, BARMM, and Caraga exhibit lower private school enrollments, with a greater dependence on public schools, aligning with the economic reality that many families in these regions cannot afford private school tuition.

A significant drop in enrollment across all sectors around 2014-2015 suggests potential factors such as policy changes, funding issues, or external influences impacting student enrollment. However, after 2016-2017, enrollment numbers appear to stabilize, possibly due to improved education plans and initiatives aimed at encouraging more students to enroll. On the other hand, the decline with the SUCs or LUCs sector has partially recovered, but it continues to remain below the private and public sectors in terms of enrollment numbers. To address these, it is important to enhance public education to ensure that it remains a competitive option especially for the middle-income families. With the private sector, it is significant to increase access in rural areas offering financial assistance programs and scholarships which can engage development in these areas. Nonetheless, there should be support for the SUCs or LUCs growth for its competitiveness with the public and private sectors.

B.3. Grade Levels

The data indicates a significant drop in enrollment as students transition from elementary to secondary levels, particularly in Grades 9 and 10. This suggests potential retention challenges at the secondary level. Additionally, non-grade primary and

secondary programs have the lowest enrollment numbers, highlighting a need for further analysis of their effectiveness and alignment with overall educational goals. To address these issues, schools should implement targeted interventions such as academic support programs, counseling services, and career guidance to improve student retention. Additionally, a review of non-grade programs is necessary to ensure their relevance and impact on student outcomes.

B.4. Gender

The data indicates that there are slightly more male enrollees than female enrollees across most regions, though the gender distribution remains relatively balanced overall. This suggests a positive trend toward gender equality in education in the Philippines, with both genders participating in the educational system at similar levels. The near-equilibrium between male and female enrollment is a sign of progress in ensuring equal access to education for all students, regardless of gender. For actionable recommendation, it is essential to continue promoting gender equality where there are equal opportunities for education. Another one is to support female education initiatives where it encourages female participation in education in particular fields where women are underrepresented.

B.5. Clusters

Clusters are generated using K-means, an unsupervised machine learning model that groups data points based on their similarities. The results of the K-means clustering process revealed the underlying patterns of the data, providing the following discoveries on each clusters:

B.5.a. Cluster 0: *Public/Private Schools Public/Private Schools mostly on CAR, CARAGA and BARMM*

Educational access in the northern and southern parts of the Philippines (CAR, CARAGA, and BARMM) are generally favorable, with both Public and Private schools offering adequate opportunities for students. Given the average enrollment count, this cluster likely represents regions with moderately developed education systems, where the balance between the availability of schools and the demand for education is relatively stable.

B.5.b. Cluster 1: *SUCs and LUCs Schools mostly on Region 4A, NCR, and Region 3*

Compared to the first cluster, the enrollment in State Universities and Colleges (SUCs) and Local Universities and Colleges (LUCs) in various regions in Luzon (Region3), Visayas (Region 8), and Mindanao (BARMM) are relatively low. The low enrollment count imply that part of each island group with SUCs or LUCs are not the target of the elementary and junior high school students. While this is not a sign of significant problems since students still in primary and education are less likely to enroll in SUCs or LUCs, it is still important to assess other factors that may contribute to their decision, specifically if there are barriers preventing students from enrolling.

B.5.c. Cluster 2: *Public Schools mostly on Region 3, Region 4a, and Region 5*

Enrollment count in Public schools across Central Luzon, Calabarzon, and Bicol ranks the highest compared to other clusters. The public schools in these regions are characterized by significant student enrollment, reflecting strong demand for public education. This high enrollment can be attributed to factors such as

proximity to urban centers and population growth. A shared characteristic among these regions is their strong agricultural foundation. As many families depend on farming, this may explain why children tend to stay in their local communities and pursue their education at nearby public schools. While the high enrollment in public schools in these regions is a positive development, it is also crucial to monitor potential challenges, such as overcrowding. Additionally, it would be valuable to assess whether students in these public schools are still continuing their education at the tertiary level.

C. Recommendations

C.1. Allocate Resources and Improve Economic Development for some Rural Regions such as CAR, CARAGA

As shown in Figures 3.1 and Figure 3.4, the total number of enrollees from the school year 2010-2011 to 2020-2021 across all regions in the Philippines is uneven. The significant difference highlights the need for government institutions to allocate more resources and enhance educational infrastructure in the affected regions, particularly CAR, CARAGA, and BARMM. To address this, the government can prioritize these regions in the education budget by establishing more schools, especially in remote and underserved areas, and launch campaigns highlighting the long-term benefits of education to foster community participation. Moreover, considering the dense population in NCR, it is crucial to promote regional economic development to ensure balanced growth between urban and rural areas, thereby creating more opportunities in rural regions.

C.2. Focus on Improving Public School Infrastructure Since There is a Significant Amount of Enrollees in Public Sectors

Given the substantial number of student enrollees in the public sector, as indicated in Figure 3.2 and Figure 3.3, it is important for the government to focus on improving public education to ensure that students continue to receive a high-quality learning experience. Public schools often face challenges such as overcrowded classrooms, insufficient resources, and outdated facilities. These issues can hinder the ability to deliver quality education, affecting both students and teachers. To address this, the government must prioritize investments in public school infrastructure, which includes constructing new classrooms and upgrading existing ones to accommodate the increasing number of enrollees. By focusing on improving the public sector, the government can create an educational system that ensures all students, regardless of their background or location, have access to quality education.

C.3. Make State Universities and Local Colleges More Accessible to Filipino Masses

As seen on Figure 3.1, Cluster 1 has the lowest average enrollment (151.06). There are different factors that contribute to this and one of which is the limited accessibility to state colleges and universities. By focusing on scholarship programs, transportation options, and distance learning initiatives, more students will be able to enroll in these schools given that they pass the entrance exams. Additionally, the lack of public awareness on the opportunities in SUCs or LUCs may also impact the low number of students. By creating awareness campaigns and letting campus visits, more students will be interested in studying in these schools.

References

- Bai, N. (2023). Educational Challenges in the Philippines. Broken Chalk. Retrieved from https://pidswebs.pids.gov.ph/CDN/document/1691385120_64d07d20eab5b.pdf
- Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Deepchecks (2024, May 27). What is Root Mean Square Error? Calculation & Importance. <https://www.deepchecks.com/glossary/root-mean-square-error/>
- Department of Education. (n.d.). Vision, mission, core values, and mandate. Retrieved from <https://www.deped.gov.ph/about-deped/vision-mission-core-values-and-mandate/>
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- World Bank. (2020, July 15). Remote Learning, EdTech & COVID-19. World Bank; World Bank Group. <https://www.worldbank.org/en/topic/edutech/brief/edtech-covid-19>

Appendices

A. Methods: Standardizing Columns

```
Column: REGION
Current dtype: object
Converted dtype: category

elem["REGION"] = elem["REGION"].astype("category")
jhs["REGION"] = jhs["REGION"].astype("category")

print(f'Regions changed to {elem["REGION"].dtype} and {jhs["REGION"].dtype} type')

Regions changed to category and category type
```

Standardizing Columns: Region

```
Column: GRADE_LEVEL
Current dtype: object
Converted dtype: category

elem["GRADE_LEVEL"] = elem["GRADE_LEVEL"].astype("category")
jhs["GRADE_LEVEL"] = jhs["GRADE_LEVEL"].astype("category")

print(f'Grade Levels changed to {elem["GRADE_LEVEL"].dtype} and {jhs["GRADE_LEVEL"].dtype} type')

Grade Levels changed to category and category type
```

Standardizing Columns: Grade Level

```
Column: GENDER
Current dtype: object
Converted dtype: category

elem["GENDER"] = elem["GENDER"].astype("category")
jhs["GENDER"] = jhs["GENDER"].astype("category")

print(f'Genders changed to {elem["GENDER"].dtype} and {jhs["GENDER"].dtype} type')

Genders changed to category and category type
```

Standardizing Columns: Gender

B. Relevant Code Outputs

B.1. Dataset Shapes

```
elem.shape
(8568, 6)

jhs.shape
(5202, 6)
```

Dataset Shapes: Elementary and Junior High School

B.2. Dataset Information

```
elem.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8568 entries, 0 to 8567
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   SCHOOL YEAR      8568 non-null  object
1   SECTOR           8568 non-null  object
2   REGION           8568 non-null  object
3   GRADE LEVEL      8568 non-null  object
4   GENDER           8568 non-null  object
5   NUMBER OF ENROLLEES 8568 non-null  object
dtypes: object(6)
memory usage: 401.8+ KB

jhs.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5202 entries, 0 to 5201
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   SCHOOL YEAR      5202 non-null  object
1   SECTOR           5202 non-null  object
2   REGION           5202 non-null  object
3   GRADE LEVEL      5202 non-null  object
4   GENDER           5202 non-null  object
5   NUMBER OF ENROLLEES 5202 non-null  object
dtypes: object(6)
memory usage: 244.0+ KB
```

Dataset Information: Elementary and Junior High School

B.3. Dataset Describe

elem.describe()						
	SCHOOL YEAR	SECTOR	REGION	GRADE LEVEL	GENDER	NUMBER OF ENROLLEES
count	8568	8568	8568	8568	8568	8568
unique	11	3	34	8	2	5202
top	2014-2015	PUBLIC	REGION 1	KINDERGARTEN	MALE	-
freq	816	2856	456	1122	4284	564

jhs.describe()						
	SCHOOL YEAR	SECTOR	REGION	GRADE LEVEL	GENDER	NUMBER OF ENROLLEES
count	5202	5202	5202	5202	5202	5202
unique	11	3	17	5	3	3690
top	2014-2015	PUBLIC	REGION 1	GRADE 7	FEMALE	-
freq	510	1734	306	1122	2601	259

Dataset Describe: Elementary and Junior High School