

**CSC105M - BIGDATA**

Final Project Documentation

|  |  |
| --- | --- |
| **Section** | G01 |
| **Team Members** | Aquino, Kurt Neil  Choy, Matthew Seaver  Matias, Angelo |
| **Date Submitted** | August 25, 2016 |

1. Data Set

Name: Philippine's Public Schools Enrollment Dataset for S.Y. 2015-2016

Size: Total of 62.7MB for 5 Excel files

Source: <http://www.deped.gov.ph/datasets>

The data sets being used for the group’s final project were gathered from the Department of Education’s main website. DepEd provides various data sets containing information about school maintenance and operating expenses, the master list of teachers, details about water and electricity supply, and much more, of public schools across the Philippines. For this project, the group focused on using the enrollment data of primary and secondary public schools from before and after the implementation of the K to 12 program into the curriculum.

To further explain the details of the data sets to be used, the “Public Schools Enrollment Dataset” is a yearly collection of records which contains the number of male and female enrollees in each grade level, which includes special education, for all of the public primary and secondary schools within the country starting from the school year of 2011-2012 up to 2015-2016 as of September 30, 2015. The provided data comes from the compiled reports by the school heads in the enhanced basic education system under the Department of Education. All of the data sets (for primary and secondary schools) share the same general details for identifying each respective school as shown in *Table 1.1.*

|  |  |
| --- | --- |
| Field Name | Description |
| School ID | Unique 6-digit reference number issued to schools in the basic education system |
| School Name | Official name as registered in DepEd |
| Region | Geographic location of school defined according to NSCB’s National Standard Geographic Classification Code |
| Province | Geographic location of school defined according to NSCB’s National Standard Geographic Classification Code |
| Municipality | Geographic location of school defined according to NSCB’s National Standard Geographic Classification Code |
| Division | DepEd division (province or city) supervising the school |
| District | DepEd district supervising the school |

Table 1.1. General School Details

Each respective school level (e.g. primary and secondary), however, has varying enrollment data before and after the implementation of the K to 12 Program into the curriculum. Prior to K to 12, which ended on the school year 2013-2014, enrollment details from kindergarten up to 6th grade were recorded for primary school, and from 1st year up to 4th year for the secondary. After the curriculum shift, which started from 2014-2015 onwards, the enrollment details from 7th grade to 10th grade were recorded for junior high school and from 11th to 12th grade for senior high school. Additional details such as the number of enrollees for the special education programs, were also recorded for both primary and secondary school. All of the enrollee counts were separated into males and females.

For this project, the group used the datasets provided and extracted as well as combined specific data into a new model, which supports online analytical processing (OLAP). The group also developed a simple java application which allows the user to extract data from the created model as well as demonstrate various OLAP queries such as roll-ups and drilldowns. The application focuses on analyzing and displaying various enrollment details such as expected incoming enrollees for certain grade levels as well as the percentage of the projected number of students who proceed to the next grade given the current number of students.

The data model used in this project may be used by various local government or non-government organizations, or even by DepEd themselves, which focuses on gathering and analyzing various educational details such as enrollment count and enrollment trends.

1. Data Preprocessing
   1. ETL Tool and Code Description

The group decided to create their own ETL tool using by using both Java and MySQL. Instead of an actual ETL tool provided by MySQL, the group created a transformation and loading script instead, which SELECTs and transforms the required data from the data sets and then INSERTs them the group’s created schema which consists all of data necessary by the application.

Considering that the provided data sets are separated per school year, the group decided to combine all of the data sets into a single model after performing some cleaning and transforming on the provided data from the data sets.

Prior to extracting data from the data sets, the group needed to first transform the data sets into something usable on a Java platform. Given that the data sets were provided as Excel files (.xlsx), the group first transformed the files into their corresponding comma-separated value (.csv) formats in order to feed them into the group’s Java based extraction tool.

As for the cleaning of the data, the data sets provided were already in a structured form so the group did face any significant difficulties in extracting the only necessary data as no outliers were present. The main difficulty however is with the designing of the new model which will contain all of the data from all of the data sets. The group has noticed that the “District” field as mentioned in *Table 1.1.* is only present from the primary school data sets and not in the secondary schools’. Another inconsistency with the data sets is that the details regarding the enrollment counts for the special education programs are only listed in the data sets after the curriculum shift (from 2013-2014 onwards). And finally, the most significant change in format from the secondary school data sets before and after the curriculum shift as the number of the grade levels and their corresponding description in order to accommodate the K to 12 program.

Another difficulty that the group has faced is with the loading of the data from the original data sets into the new combined model due to the sheer volume of the provided data. Each data set contained about an average of 38,000 rows, and with a total of 5 data sets to be combined, it took some time to completely dump all of the data into the new model. Minor troubleshooting was also performed as the original data sets contained symbols that were interfering with the extraction and transforming process.

* 1. Database Model

During the extracting process, the group has selected only the necessary data to be collected from the provided data sets. As the original data sets contained redundant data such as the total number of enrollees per grade level despite already having separate male and female enrollee counts and even a grand total of enrollees for each specific school, the group decided to not include them as they would only contribute to the overall volume of the new model as well as considering the fact that these data can be extracted and processed from the already provided data. Considering the inconsistencies as mentioned in *Section 2.1.* the group decided to separate the models for the primary and the secondary school levels as well as their respective data before and after the implementation of the K to 12 program in order to perform an even more specific level of analytics and visualization. The data included in the newly transformed and combined data set can be viewed in *Figure 2.1.*

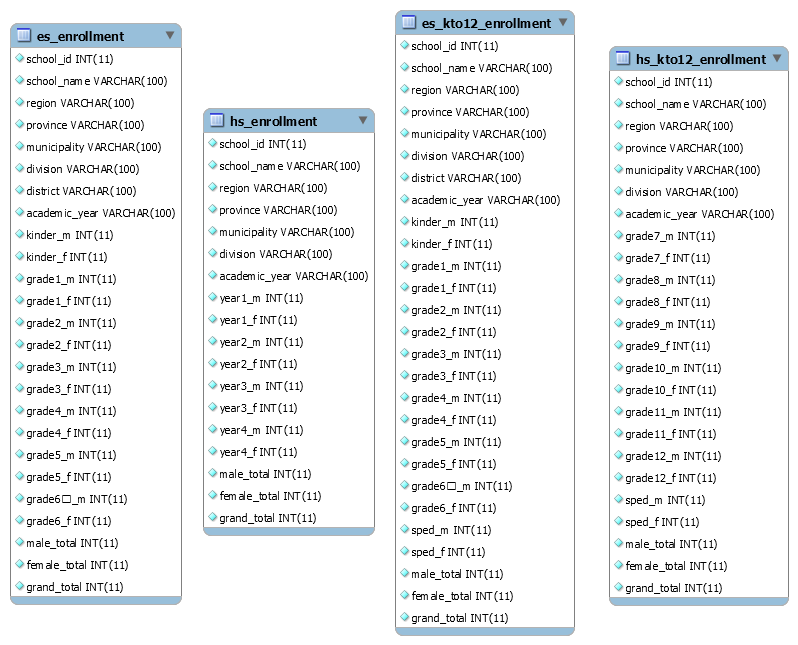


Figure 2.1. Entity Relationship Diagram

Considering that the original data sets did not have any data references as all of the data relationships are one to one (the details of a school belongs to that school alone), the group decided to imitate the format as well since adding any other entities would be unnecessary.

1. Feature Selection

Listed in this section are the features and functionalities that the group’s application aims to provide:

* 1. Expected number of incoming 1st grade elementary and 1st year high school students

This feature gives a prediction of the number of incoming 1st grade elementary students based on the trends of enrollments from the previous years of kindergarten to 1st grade transitions. As an example, given the initial school year of 2011-2012, there were 255 enrolled kindergarten students enrolled on a specific school. On the following year (2012) there has been a turnout of 243 1st grade students on the same school. While also considering the results on the succeeding years up till the school year of 2015-2016, the application will predict based on the training data what will be the upcoming turnout of the number of 1st grade enrollees for the school year of 2016-2017. The same scenario goes for the incoming 1st year high school students based on graduating 6th grade students from the previous years.

This feature will be mainly using the enrollment data from the kindergarten, 1st Grade, 6th grade, and 1st year high school fields.

* 1. Enrollment trends per grade level per year

This feature presents the enrollments per grade level per year by presenting them as a histogram or as line chart. The data to be presented may be sorted by school, by Region, by Province, by Municipality, by Division, by District, or by any combination of the listed locations. This allows for the comparison of enrollment data by different schools or regions for analysis purposes. For example, a certain Province has a significantly lower count of enrolled students as compared to another, considering other factors (which are not included in this application) such as employment rates as well as the average annual income of the people residing the province, this may give an idea on why so few students are enrolled. Again, this is just a sample scenario and this feature will only provide the visualization of data.

This feature will make use of the enrollment data from all of the grade levels while synchronizing the data from each school’s yearly reports.

* 1. Percentage of public schools that offer a special education program

This feature presents the visualization of the number of schools which provides the special education program within their curriculum we well as the percentage of students enrolled in the special education program versus those who are not per school. This information provides an idea which public schools mainly focus on special education in order to determine which public schools may need specialized facilities to provide for the needs of special children.

This feature will focus on the number of enrolled students in the special education programs of both primary and secondary schools as well as the total number of regular students.

1. Visualization

Majority of the visualizations developed for the application were generated using Java’s JavaFX graphical chart library. Listed here are the visualizations used in order to present the data:

* 1. Expected number of incoming 1st grade elementary and 1st year high school students

Considering that only a prediction will be provided for this feature, the data will only be presented in a tabular format, with the expected number of enrollees for each school being the main focus. The data being presented may be further drilled down or rolled up depending on selected specific locations in order to trim down the result set. As this feature deals with the enrollment trends from kindergarten to 1st grade and from 6th grade to 1st year high school, the visualization may be viewed from *Section 4.2.*

* 1. Enrollment trends per grade level per year

As mentioned in *Section 3.2.,* this feature presents the enrollments per grade level per year by presenting them as a histogram or as line chart as shown in *Figure 4.1.* The data to be presented may be sorted by school, by Region, by Province, by Municipality, by Division, by District, or by any combination of the listed locations. This allows for the comparison of enrollment data by different schools or regions for analysis purposes. In this example, the enrollment data of a certain school is presented in order to see the school’s enrollment trends throughout the course of 5 years (SY 2011-2012 to SY 2015-2016).

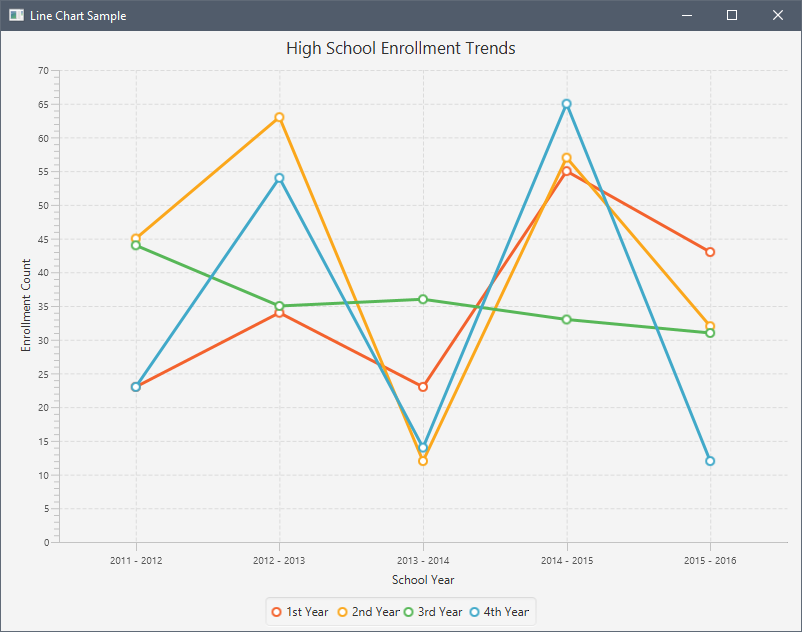


Figure 4.1. Sample Enrollment Trend

* 1. Percentage of public schools that offer a special education program

As this feature presents the visualization of the number of schools which provides the special education program within their curriculum as percentages of students enrolled in the special education program versus those who are not, a pie chart will be used in order to display the data as shown in *Figure 4.2*. This information provides an idea which public schools mainly focus on special education and which schools does not have special education students in order to determine which public schools may need specialized facilities to provide for the needs of special children. In this example it can be observed that only a small percentage of students are enrolled in the special education program from the schools across Region 1.

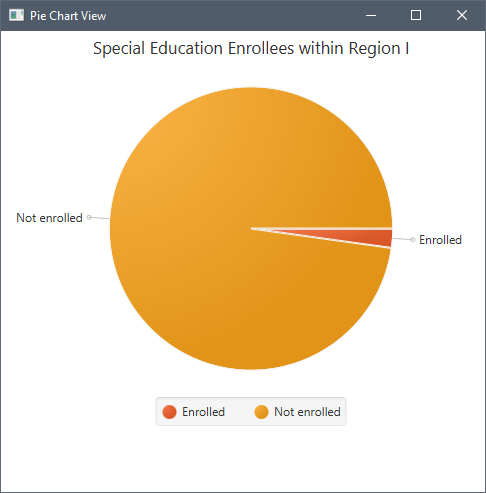


Figure 4.2. Sample Pie Chart View

1. Analytics
   1. Expected number of incoming 1st grade elementary and 1st year high school students

In order to formulate the predictions for the incoming grade levels, the group attempted to formulate a neural network by using previous enrollment data from the kinder and 6th grade enrollees as inputs and from the 1st grade and 1st year high school enrollees as outputs for the training data. Considering that Java has no local support or a library for supporting neural network functionalities, the group was unable to directly link the prediction feature to the main application as an API was used. The API used was created by Kevin Amaral from the University of Massachusetts Boston, which is named as “WekaAPI”. This API ports features and functionalities from the Weka software which is a collection of machine learning algorithms for data mining tasks.

* 1. Enrollment trends per grade level per year

This feature did not require any analytics as it merely presents the provided data from the data sets. The group did however, allow for the drill down and rollup of data in order to present generalized or more specific data depending on the specified location.

* 1. Percentage of public schools that offer a special education program

In order to get the percentage of public schools that offer a special education program, the group simply got the total number of students enrolled in a special education vs the total number of students who are not which did not require a special algorithm to calculate.

1. Interpretations, Findings, and Conclusions