CSIT 356: Introduction to Data Science
Project Report

Project Title: Data Research on Enrollment by School Year (Schools) From 2017 to 2018 in the State of Massachusetts

Team Members: Edwin Asitimbay, Daniela Bautista, John Lee

---

1. Description

1.1 Basic Information
We want to see the amount of enrollments, in both the highest and lowest out of all the schools in the Massachusetts state during the 2017-2018 academic year, as well as the overall average.

1.2 Project Objectives
Which school has the highest and lowest enrollment, and what is the overall average?

1.3 Description of the Data Set
The data set records the data of enrollment in the state of Massachusetts during the 2017-2018 academic year. It highlights each city (municipality) and each school within a city. The data frame consists of not just the total number of enrollments, but also the number of enrollments base on grade level, special classes, nationality/race/sex, language profienciency, disability, income standing, and high needs. Along with numbers, the data set also captured the percentage. However, despite having the columns related to income standing, no data has been collected on those.

2. Exploration of Data Analysis

2.1 Data Preparation
Before we could load the data, we had to import pandas in order to use the read_csv() function. After that, we have to put the correct directory and name of the csv file into the read_csv() function to load in the data and make it into a dataframe.

As for data cleaning, we first renamed all of the columns due to them being unclear on what they represent. For example, one column is named “swd_num”, which the “swd” could mean a number of things. From the resource website, it actually represents the Students With Disabilities enrolled in each school. Therefore, to make the names more clear and detailed, we renamed all of

the columns with the df.rename(columns = {...}) function and then assigned the changes to the dataframe. Next, we removed the columns that collected the data in percentage, as they do not serve anything useful to us while the answers to our objective are collected in numbers, which is more useful for being very specific. Same goes for the columns that are only filled with NaN, for it means these data are missing or never collected, so they serve no purpose for us. Lastly, there were two columns called “ed_num” and “ed_pct”, which the resource website provides no detail on what they are. Because we cannot determine what they are, and they do not harm the data set if manipulated, they were removed. To remove these unnecessary data, we used the dp.drop([...], axis = 1) and assigned the changes to the dataframe.

Key Codes:

```
import pandas as pd
df = pd.read_csv('tabular.educ_enrollment_by_year_schools_2017-18.csv')

dp.rename(columns = {...}, inplace = True)
dp.drop([...], axis = 1, inplace = True)
```

2.2 Data Analysis using Descriptive Statistics

In order to calculate the statistical measurements from the file, the total of every column needed to be removed. If not removed then it would throw off the calculations and lead to inaccurate results. We kept the original dataframe and made the changes in df1 for the calculations. The Average amount of enrollments was around 515 students for every school for 2017-2018 school year. It ranges from pre-kindergarten to grade 12th. Noting this every school in Massachusetts does not have all 14 grades in one school so their specific enrollments would be different. The Max amount of students enrolled in one school this year was 4,123 students in Brocton High. The minimum number of students enrolled was 1 and it comes from Fall River Gateway to College @ BCC, Landmark School and Solstice School. This can be due to low income, not enough space or supplies.The standard deviation of the amount of enrolled students in 2017-2018 was about 370. Since it's such a high number that means the data is very much spread out. Meaning it fluctuates from a high to low number. Meaning depending on which grade the school is able to teach. The Higher the grade the more they are able to enroll into the school and the lower the grade it necessarily doesn’t mean you can enrol as much but you will need a student to teacher ratio that works really well.

Key Codes:

```
df1 = df
df1 = df1.drop('State Totals')

Mean = df1[...].mean()
Median = df1[...].median()
```

Dmax = df1[...].max()
Dmin = df1[...].min()
Std = df1[...].std()

df1[df1['Enrolled'] == 1]
df1[df1['Enrolled'] == 4123]

2.3 Other Techniques

Technically not data cleaning nor a technique, but we manipulate the data set by organizing it further, setting the index and sorting the data in ascending order through the column “School Name”, using df.set_index() and df.sort_values(). This way, the dataframe would be much easier to read. These changes are then assigned to the dataframe.

Key Codes:
df.set_index('School Name', inplace = True)
df.sort_values('School Name', inplace = True)

3. Data Visualization

Since there were almost two thousands rows of schools, we decided to choose schools from a specific municipal/city. We choose the municipal Acton, that will give us the number of students that had enrolled in any schools in Acton. After we got the schools and the number of students enrolled in those schools in Acton, we can finally create Data Visualization. The first thing is using plt.ylabel to label the y-axis as the number of students and using plot.bar() to create the bar graph and a title for the bar graph titled “Students Enrolled in Schools in Acton, Massachusetts of 2017 - 2018.” We also included rot so the labels on the x-axis can be read properly without intersecting each other.

Key Codes:
df2 = df1[df1[‘Municipal’] == ‘Acton’] [“Enrolled”]
df2

df2.plot.bar(title = “...”, rot = 70)
plt.ylabel[“Number of Students”]

4. Conclusion

We can conclude that our objective was meant in answering which school has the highest and lowest enrollment, along with what was the average enrollment. However, we did not expect our result to have more than one school of lowest enrollment, nor did we expect to have no school with absolute 0 enrollment. It also seems that the schools who teach a higher grade such as High

Schools tend to have a much higher enrollment rate since there are more subjects/classes to be taught and attended.