Capstone Project

**Introduction**

1. Background and Problem

In today’s job market, a college degree is essential for many high-paying jobs. There are many articles and studies that show the relationship between salary and educational attainment. For example, according to an article published by Northeastern University, the average salary for those with high school diploma is $38,792, compared to $64,896 for those with bachelor’s degree[[1]](#footnote-1). However, there are many high school students who have been struggling to keep their grades high enough to enter colleges.

Based on an assumption that grades in high school are positively related to college entrance, this project examined relationships between academic scores (e.g., ACT) and college enrollment rates in high schools in Chicago. This project also tried to identify nearby universities who could possibly provide educational services (e.g., mentoring or tutoring) to students in high schools who have the lowest academic scores in Chicago.

1. Interest

The results of this project may be utilized by high school teachers, parents, non-profit organizations, universities, and other community members to develop plans to provide educational services to high school students who have been struggling to keep up their classes.

**Data**

1. Data Sources

Data that is provided by Chicago Public Schools was used. The dataset is available from the website below:

https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t?cm\_mmc=Email\_Newsletter-\_-Developer\_Ed%2BTech-\_-WW\_WW-\_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm\_mmca1=000026UJ&cm\_mmca2=10006555&cm\_mmca3=M12345678&cvosrc=email.Newsletter.M12345678&cvo\_campaign=000026UJ

Due to limited availability of the most current datasets, the data that was published in “Progress Report Cards (2011-2012)” was used. Variables used for this project are:

* 9th Grade EXPLORE (2010): average score of EXPLORE test
* 10th Grade PLAN (2010): average score of PLAN test
* 11th Grade Average ACT (2011)
* College Eligibility %
* Graduation Rate %
* College Enrollment Rate %

Scores of EXPLORE and PLAN tests in 2011 were not available in this dataset. Therefore, the ones in 2010 was used. There were multiple scores that may be specific to the 12th grade students, but no detailed descriptions were available for those scores, so academic scores of the 12th grade students were not included.

Since this project also tried to identify locations of universities in Chicago, geospatial data was obtained from Foursquare.

1. Data Cleaning

There were 93 high schools in this dataset. Among these high schools, there were 18 schools who did not have the complete data for this project. Using Nupmy, the schools with incomplete data were dropped.

**Methodology**

1. Descriptive Statistical Analysis

Descriptive statistical analysis was conducted to see some basic information of the dataset.

Table

Description automatically generated

Figure1: Descriptive statistical analysis

Among the 75 high schools in Chicago, mean scores for 9th, 10th, and 11th graders are 14.1, 15.12, and 16.85 respectively (the highest possible points for those tests are 35, 25, and 36). Mean graduation rate is 63. Since the graduation rate in IL in 2010-2011 is 83.8[[2]](#footnote-2), the rate in Chicago is approximately 20 percentage points lower than the state’s rate.

1. Regression Plot

By using Seaborn’s regplot function, regression plots were generated to see the relationships between the independent variables and dependent variable.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure 2: Regression plots

There regression plot figures show that each academic score has a positive relationship with the college enrollment rate.

1. Multiple Regression Analysis

In order to examine how much influence those academic scores have on the college enrollment rate, multiple regression analysis was conducted, and scikit-learn’s LinerRegression function was used. Then, the R-square score was calculated. The R-square was 0.67, meaning that the model generated by using those academic scores explains 67% of the variability of the college enrollment rate.

1. K-Means Clustering

Next, those 75 high schools were divided into four groups by using K-Means Clustering, based on those academic scores.

A picture containing application

Description automatically generated

Figure 3: Numbers of high schools in four groups

The figure above shows how many high schools are in each group. Group 0 has the most schools (35) and Group 3 has the least (5). Then, the average values of each variable were calculated.

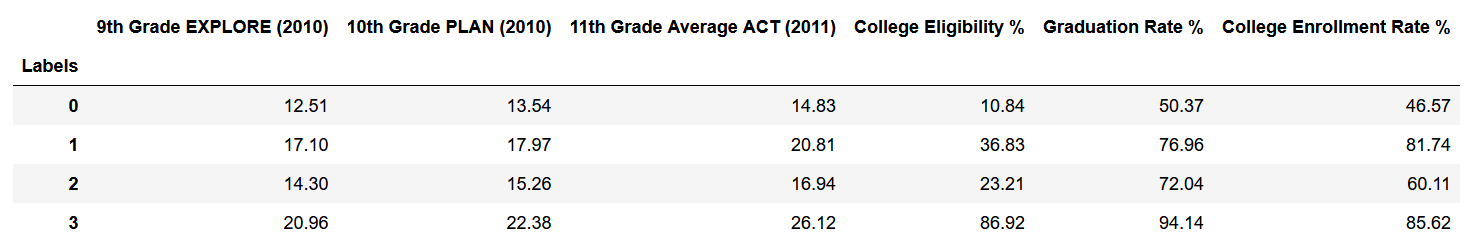


Figure 4: Mean values of academic scores and college-related records of the four groups

The figure above indicates that the Group 0 has the lowest scores in all variables. That means this group is the one who needs educational services from universities the most. In order to visualize Figure 4, bar charts were created by using matplotlib.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Figure 5: Bar charts of the academic scores and college-related records of the four groups

1. Geospatial Data from Foursquare

Lastly, geospatial data of universities was obtained from Foursquare. Then by using folium, locations of universities were plotted on a map to see if there are any universities that are located around the high schools in Group 0. If there are, students from those universities could help those high school students by providing tutoring or mentoring services.

In order to get a general idea of how universities are located in Chicago, the geospatial data was retrieved from Foursquare, which contains the first 100 universities (radius = 20,000 meters) in Chicago. Duplicates in this dataset (i.e., universities that have exactly the same names) were deleted. In total, 91 universities’ location data was obtained.

First, the locations of high schools were plotted as below.

Map

Description automatically generated

Figure 6: Locations of high schools in Group 0

Then, the locations of the universities (red dots) were plotted on the same map.

Map

Description automatically generated

Figure 7: Locations of universities in Group 0.

**Results**

From all the analysis that were conducted in this project, it is clear that the academic scores in high school and college enrollment rate were positively related. The independent variables that were used in the multiple regression analysis explain more than half of the variance of the dependent variable. The results indicate that those who are at risk of having low academic scores or grades in high school need interventions or educational services to increase the chance of being admitted to colleges.

One possible solution is to find nearby universities and seek any collaboration opportunities with them. The maps show that there are a lot of universities in the center and northern part of Chicago, meaning that it might be relatively easier for those high schools located in these areas to seek the collaboration opportunities. However, there are fewer or no universities located in the southern part of Chicago. That means that it might be harder for those high schools in that area to seek the collaboration opportunities with the universities.

**Discussion**

This project shows some challenges faced by specific high schools, especially those in the southern part of Chicago due to limited possible opportunities to work with the universities to improve the students’ academic experience. However, since there are many universities in Chicago, it is still possible to work with them. For example, the universities could identify students who live close to those high schools, and also have a strong passion to help those students. In return, these universities could offer their students some credits that are necessary to graduate.

Another possibility is to have online tutoring or mentoring services through zoom or other video-chat apps. During this pandemic time, video chat apps have become more common, so it might be easier than ever to implement online tutoring services.

Finally, if collaborating with universities is difficult for some schools, seeking assistance from local communities or non-profit organizations might be possible. Since this project identified high schools with low academic scores, it is not difficult to target those who have been struggling to keep up with their classes.

**Conclusion**

In this project, the relationship between the academic scores and college enrollment rate was examined, and then high schools with low academic scores were identified. In addition, this project plotted those high schools and the universities to see how close these universities are located. Even though some challenges and difficulties faced by specific high schools were shown, there are some possible solutions that could be discussed. It is important to note that data analysis is not only to identify problems but also to suggest solutions for the future.

1. <https://www.northeastern.edu/bachelors-completion/news/average-salary-by-education-level/> [↑](#footnote-ref-1)
2. <https://datacenter.kidscount.org/data/tables/9782-high-school-graduation-rates#detailed/2/any/false/1696,1648,1603,1539,1381,1246,1124,1021,909/any/19056> [↑](#footnote-ref-2)