﻿**Final Project-Summary**

Yihao (Henry) Li, Weixuan (Tony) Jia

**Introduction:**

This project focuses on analyzing whether high schools in the Chicago Public Schools system with higher proportion of non-White students (especially African American and Hispanic students) and different locations are facing problems on getting students prepared for college education. Specifically, we use the “College Readiness Index”, an indicator which depicts 12th graders participation in the Advanced Placement and IB tests, from the U.S.News website as the measure of academic performance, and draw connection with the ethnicity of various high schools across the city to see their relationship. Moreover, we bring the factor of location into the analysis as well to see if locational differences (i.e. schools with similar higher proportion of ethnic minority students but at different places) play a role in the academic preparation for college education.

**Structure:**

We applied different techniques to analyze the research questions. For the locational differences, we used data visualization tools to show the distribution of public high schools across the city and the corresponding College Readiness Index; For the racial part, we chose a simple OLS model and machine learning models to identify whether ethnic minorities are exposed to less academic resources regarding college preparation. The detailed structure is in the pdf file *structure.pdf.*

**Data Visualization**

We first provide two plots (in html format) to show the overall distribution of the Chicago Public High Schools (please see the heat\_map.html and school\_map.html for more info). It’s clear that the number of schools is big on the west side of the downtown area and small at the southern area. However, the population density is relatively high in the southern part of the city. Therefore, there exists a mismatch in the quantitative perspective between population and the number of public high schools.

The four subplots *race\_vs\_college.png*, which are intended to show the relationship between College Readiness Index and share of different races, present a vague correlation between a school's share of African American students and how well a school prepares its student to be ready for college. Further investigation is demanded.

Chart, scatter chart

Description automatically generated

Regarding the locational difference, we can observe from the two plots below *(College\_Readiness\_Geo .png and num\_map.html)*. Notice that the left graph depicts the no.1 ethnic group in each high school (black dots for African American students, blue for Hispanic, and red for White) and the right depicts the average College Readiness Index scores across different districts. From the left graph, we can see that African American and Hispanic student are the dominant ethnic group. As the choropleth of College Readiness Index by Chicago neighborhood, *College\_Readiness\_Geo.png*, shows, the dark-colored neighborhoods, meaning neighborhoods having schools with low performance on preparing students for college, spread over Chicago. It is worth mentioning that for those districts which have high average Index scores, the number of high schools with Hispanic to be the dominant ethnic group is higher than that of with African American students. This is somewhat coincidental with the plots we have above. We will prove (partly) whether this is true in the next part (that African American students in public high schools are receiving less college prep assistance).

Map

Description automatically generated Chart

Description automatically generated

**OLS model:** A simple OLS regression is given by regressing different ethnic group (i.e. the proportion of African American students) on the College Readiness Index. We can see that the R-square for this result is not high (from the OLS regression results picture below). However, we can still point out some interesting result, including the negative sign in front of the coefficient for African American students (the coefficients of the Pct independent variable of other ethnic groups are all positive), indicating a possible relationship that one unit/percentage increase of African American students in the Chicago Public High School is related to about 0.23 decrease of the College Readiness Index. This result further provides the intuition regarding the unequal educational resources available for ethnic minorities (especially African Americans in this case) within in the Public School system.

A picture containing text, plaque, screenshot

Description automatically generated

**﻿**In order to discover the above relationship in a more solid manner, we then include a separate model to incorporate the types of schools (neighborhood, charter, magnet etc.) as a dummy variable. According the principle component analysis we have conducted, only including the percentage of white and African American students is sufficient to find the relationship between other student ethnicities, such as Asian and Hispanic, and a school's performance on preparing its students being ready for college. After we put control on school types in the regression, observations generated from the regression without control on school types still hold. A school's higher percentage of African American students is associated with a decrease in College Readiness Index, and a higher percentage of White students is associated with an increase of College Readiness Index. For school types, being a magnet school, a school located in neighborhoods, or a contract school also associate with a decrease in College Readiness Index.

**A picture containing graphical user interface

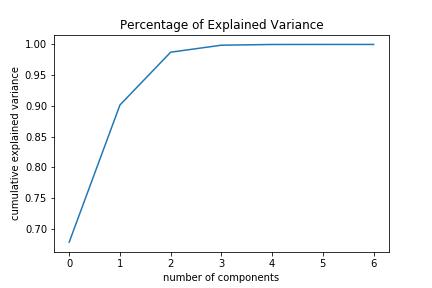
Description automatically generated**Text

Description automatically generated

**Machine Learning Model**

﻿Since we decide to find out whether a school's student racial proportion is associated with how well each school prepares students for college, indicated by the value College Readiness Index, the model of principal component analysis is chosen.

﻿As the explained variance graph, *Explained\_Variance.png*, indicates, two components are sufficient to explain nearly 100% of the variance of a school's student race proportion and College Readiness Index. Therefore, we decide the number of principal components is .



The loading matrix below points out that ﻿the first loading vector places a positive direction on the percentage of black students, *Pct\_black*, and the percentage of students with unknown ethnicity, *Pct\_unknown*. All other variables, including College Readiness Index, are placed in a negative direction by the first loading vector. A negative correlation between how well a school prepares its students for college is negatively associated with the number of African American and unknown-ethnicity students it

has. Such finding is exhibited by the second loading vector as well.

|  |  |  |
| --- | --- | --- |
| ﻿ Variable | PC1 | PC2 |
| Pct\_white | -0.015722 | 0.368365 |
| Pct\_black | 0.715533 | -0.217300 |
| Pct\_native | -0.002378 | 0.001122 |
| Pct\_hispanic | -0.684347 | -0.399987 |
| Pct\_asian | -0.017850 | 0.169455 |
| Pct\_unknown | 0.001564 | 0.032998 |
| College Readiness Index | -0.138227 | 0.792019 |

**Limitations:**

* **Data availability**: The limitations can be described from several perspectives. On one hand, different platforms have different names for a same high school, and it is time-consuming to compare the names. Due to the complexity of these names, it is hard to merge the datasets with ethnicity info with College Readiness Index on a one-to-one basis. As a result, we are only able to get a sample (fortunately, this is random) from all the high schools. This is shown by the blank sectors on the *College\_Readiness\_Geo.png* plot. Besides, College Readiness Index is only one measurement of the academic capability of a high school. There should be other indicators who could use to generate different results.
* **Statistics**: On the other hand, the OLS model is far from perfect. Even though we included dummy variables into the analysis, there are many other factors which could impact the result, including economic, student-teacher ratio, and even crime rates around the peripheral of each school. For example, many wealthy families would send kids to private schools to prepare for college education, and the public school system would be filled by those students from relatively poor families (a lot of these families are from the ethnic minority group) whose income would not even allow their kids to go to college. These omitted variables might provide a possible explanation for the low R-square number from the regression summary. In this sense, the result of the OLS model needs more adjustments.

**Documents/files description:**

**Boundaries - ZIP Codes:** the shapefile of Chicago neighborhoods’ geographic information

**Finalproject-usnewsData.py:** Scrap the U.S. News College Readiness Index

**Chicago\_Public\_Schools\_-\_School\_Profile\_Information\_SY1617.csv:** a raw dataset including all **Chicago high schools’ school**-type, geographic information and contact information

**College\_Readiness\_Geo.png:** a choropleth map describing College Readiness Index by neighborhood

**data\_college\_readiness.csv:** a raw dataset, generated by scrapping US News website, including Chicago high schools’ College Readiness Index

**Explained\_Variance.png:** the graph depicting the cumulative variance explained by the principal components

**Final\_Version\_dataframe.csv:** the dataframe, generated after data-cleaning, entailing school’s College Readiness Index, percentage of different student ethnicities, and geographic information

**finalproject\_by\_Yihao(Henry)\_Li.py:** the code script written by Henry Li;

**race\_vs\_college.png:** four scatter plots narrating share of student race vs. College Readiness Index

**school\_ethnicity.xlsx:** a raw dataset entailing each school’s share of student ethnicities

**heat\_map.html:** The heat map of school distribution

**num\_map.html:** The distribution of the top ethnic group in each school

**school\_map.html:** School distribution

**schooldistribution.py:** codes of plotting the above three graphs

**index\_ols:** Perform the OLS regression

**Citations:**

Geographic info of Chicago Public Schools

﻿<https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/8i6r-et8s>

U.S. News-College Readiness Index

<https://www.usnews.com/education/best-high-schools/illinois/districts/chicago-public-schools-110570>

Spatial Data-Chicago districts boundaries

﻿<https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-ZIP-Codes/gdcf-axmw>