High School Longitudinal Study

Group 2

Fan Ding
University of Minnesota, Twin Cities
ding0322@umn.edu

Rory Flemming
University of Minnesota, Twin Cities
flemm053@umn.edu

Connor Wick
University of Minnesota, Twin Cities
wickx182@umn.edu

Xianjian Xie
University of Minnesota, Twin Cities
xie00250@umn.edu

Abstract
Education outcomes may be affected by a variety of individual, social, and environmental factors. The amount of impact different situational factors have on educational outcomes is not clear. We used a large-scale longitudinal dataset including demographics and survey measures from students, parents, and their schools to investigate the degree to which factors such as poverty, parental education, and socioeconomic status affect academic outcomes. We found that poverty, parental education level, family income level, and the number of AP courses taken affect high school GPA but not school urbanicity. We arrived at conclusions aimed at increasing the average GPA scores across high schools based on our findings that academic performance is impacted by these situational variables. And also we gave up some suggestions to local governments to increase students' GPA, such as more job opportunities for parents, or support local businesses.

I. INTRODUCTION

Education in the USA has been known to underperform educational expectations with respect to the rest of the world, considering it is one of the wealthier countries. The Pew Research Center aggregates data from the Programme for International Student Assessment (PISA) which shows that the USA overall placed a mediocre 38th overall out of the total seventy-one countries [1]. Scores in the USA overall have hardly changed in the last two decades [1]. These scorings are created by collecting. It has been speculated that this result may represent the mixture of US education circumstances, where the heterogeneity of educational, economic, and social resources [2]. To investigate how such heterogeneous situational factors may affect educational outcomes, we look to a large-scale longitudinal data set that collects data on a wide range of factors and educational and occupational outcomes.

Dataset

The High School Longitudinal Study of 2009 [3,4] is a large study of high school students in an attempt to provide enough data to make conclusions about students' academic performance. The dataset consists of an initial study in 2009, and then multiple follow-ups. The features of this dataset include both academic records, self-report demographic data on individual students, their family and school, and answers to survey questions about academic and extracurricular life. Data is collected in the first, third, and fourth years of high school and then both three and four (unavailable) years after high school.

II. METHODOLOGY

a) Tools

We used MySQL Workbench, CSV, LucidChart, and Google Docs to implement our project. We used MySQL Workbench as our main development tool. We used it to input and visualize our data and implement code and output tables to show what's the factors that affect high student's GPA. MySQL Workbench is a unified visual tool for database developers. It provides data modeling, SQL development, and much more. The major reasons we chose it are it's beginner-friendly and it provides a visual console to easily visualize the data. We used LucidChart to implement our ER diagram. LucidChart is a web-based proprietary platform that allows us to collaborate on drawing, sharing chars, and diagrams. We used Google Doc to keep track of each meeting's notes. Included: weekly tasks and information links.

b) Preprocessing

The student variables and school variables were first downloaded from the HSLS09 site [4]. The student variables table, being too large to import into MySQL Workbench, was read by chunks as Pandas data frames. Then, features of interest were selected and appended to a new table. This new, smaller student table containing only features of interest was imported into MySQL Workbench for querying. For these queries, the school variables table was not needed.

c) ER model

With thousands of features in the dataset, we only focused on the ones we used to make the diagram. These involved the student and their variables, the school and its variables, and the student's family as a weak entity with its own variables.

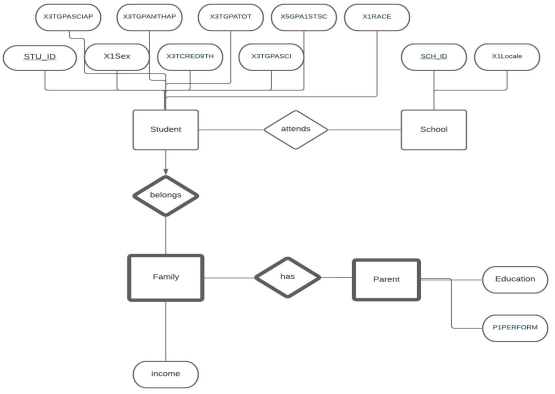

Fig 1. The ER diagram of our simplified dataset.

d) Relational table

The data, once preprocessing was completed, had 23,503 rows and 19 columns in one table. Since every query we did was on the student entity, we only had to work with one table. The primary key for student is STU_ID. The first couple of rows and columns can be seen below.

TABLE 1

STU_ID	X1SEX	X1PAR1EDU	X1MOMEDU	X1PAR2EDU	X1DADEDU	X1FAMINCOME	X3TCRED9TH
10001	1	5	5	5	5	10	8
10002	2	3	3	2	2	3	7
10003	2	7	7	-7	0	6	6
10004	2	4	0	-7	0	5	9
10005	1	4	4	-7	0	9	7
10006	2	3	3	3	3	5	8
10007	2	2	2	-7	0	4	7
10008	1	5	5	7	7	7	7.5
10009	1	2	2	2	2	4	8
10010	2	3	3	2	2	4	8
10011	1	3	5	5	3	5	8
10012	2	7	7	7	7	13	8
10013	1	7	7	7	7	13	6.5
10014	2	2	2	-7	0	2	-8

Reduced student variables table containing our variables of interest.

III. Queries and Result

1.How poverty affects high student's GPA

We used a percentage table to show a comparison of students that live in poverty and students that are not in poverty. Here is our code:

```
WITH number_of_student AS(
    SELECT  X3TGPATOT, COUNT(*) AS numbers
    FROM project.hsls_16_student_projectvars
    WHERE X1POVERTY=1
    GROUP BY X3TGPATOT
    ORDER BY X3TGPATOT DESC
),
total_student AS (
  SELECT SUM(numbers) total_student
  FROM number_of_student
)
SELECT X3TGPATOT, numbers, ROUND(numbers / total_student,3) percent
FROM number_of_student, total_student;
```

Fig 2. Code for poverty affects students GPA

TABLE 2

X3TGPATOT	numbers	percent
4	100	0.037
3.5	296	0.111
3	473	0.177
2.5	503	0.188
2	486	0.182
1.5	323	0.121
1	203	0.076
0.5	88	0.033
0.25	38	0.014
-8	159	0.060
-9	2	0.001

X3TGPATOT	numbers	percent
4	1626	0.116
3.5	3439	0.245
3	3219	0.229
2.5	2397	0.170
2	1433	0.102
1.5	673	0.048
1	305	0.022
0.5	108	0.008
0.25	33	0.002
-8	796	0.057
-9	33	0.002

GPA distribution in Poverty group and non-poverty group (Left: poverty. Right: non-poverty)

Poverty affects Student's Total GPA tremendously. Easy to see that in the poverty group, students with a GPA of 4.0 only 3.7%. However, in the non-poverty group, that number goes up to 11.6%. students in the poverty group have a grade A. Besides, 30.5% s GPA<2. But in the non-poverty group, only 13.8% of students in the non-poverty group have GPA<2. To find out the deeper reason for poverty, we implement the table below to find out the relationship.

TABLE 3

HaveJob	Poverty	numbers
0	1	205
1	1	349
1	0	3827
0	0	635

Result table for Employment affects Poverty

Around 40% of families in poverty don't have jobs. To solve the low GPA problem caused by poverty, increasing job opportunities is very necessary.

2.How family income relates to high school student's GPA

In the original data set, the family income of students is divided into 13 different levels. We group these students by different levels of family income and calculate the average GPA of students in each group. Following shows the SQL query and the corresponding output:

```
SELECT X1FAMINCOME AS family_income_level, ROUND(AVG(X3TGPATOT),3) AS avg_GPA
FROM hsls_16_student_projectvars
GROUP BY X1FAMINCOME
HAVING COUNT(*)>10
ORDER BY  X1FAMINCOME;
```

Fig 3. MySQL code for querying GPA average with respect to family income level.

TABLE 4

family_income_level	avg_GPA
1	2.278
2	2.446
3	2.684
4	2.869
5	3.013
6	3.076
7	3.117
8	3.188
9	3.152
10	3.204
11	3.264
12	3.286
13	3.303

Average GPA for each family income level. Family income levels are ascending with respect to yearly income.

TABLE 5

Category	Label
1	Family income less than or equal to $15,000
2	Family income > $15,000 and <= $35,000
3	Family income > $35,000 and <= $55,000
4	Family income > $55,000 and <= $75,000
5	Family income > $75,000 and <= $95,000
6	Family income > $95,000 and <= $115,000
7	Family income > $115,000 and <= $135,000
8	Family income > $135,000 and <= $155,000
9	Family income > $155,000 and <=$175,000
10	Family income > $175,000 and <= $195,000
11	Family income > $195,000 and <= $215,000
12	Family income > $215,000 and <= $235,000
13	Family income > $235,000
-9	Missing
-8	Unit non-response

Code labels for family income labels. Higher labels indicate higher family income.

With the increase of family income, the average GPA shows a monotonously increasing trend.

3. How school urbanicity affects student's GPA

The location of the high school the student attended was placed into one of four categories with respect to its urbanicity. From most to least urban these are the city, suburb, town, and rural. The four tables from this query are ordered as such and are shown below.

TABLE 6

X3TGPATOT	numbers	percent
4	634	0.095
3.5	1372	0.205
3	1371	0.205
2.5	1148	0.172
2	792	0.118
1.5	451	0.067
1	256	0.038
0.5	128	0.019
0.25	51	0.008
-8	471	0.070
-9	15	0.002

GPA distribution of students living in cities.

TABLE 7

X3TGPATOT	numbers	percent
4	656	0.077
3.5	1734	0.205
3	1740	0.206
2.5	1462	0.173
2	1076	0.127
1.5	581	0.069
1	302	0.036
0.5	151	0.018
0.25	58	0.007
-8	675	0.080
-9	32	0.004

GPA distribution for students in suburbs.

TABLE 8

X3TGPATOT	numbers	percent
4	262	0.094
3.5	554	0.199
3	567	0.203
2.5	484	0.174
2	384	0.138
1.5	216	0.077
1	112	0.040
0.5	42	0.015
0.25	6	0.002
-8	158	0.057
-9	3	0.001

GPA distribution for students living in towns

TABLE 9

X3TGPATOT	numbers	percent
4	493	0.089
3.5	1073	0.193
3	1179	0.212
2.5	1090	0.196
2	751	0.135
1.5	409	0.074
1	198	0.036
0.5	60	0.011
0.25	33	0.006
-8	271	0.049
-9	2	0.000

GPA distribution for students living in rural areas.

There does not seem to be a distinguishable difference between the GPAs of students who went to schools in a city or suburb when compared to a town or rural schools.

4. How parent's educational level affects student's GPA

This variable deals with the student's parents and their educational level. From top to bottom, ordered from least to most education we have neither parent has bachelor's degree or anything above, at least one parent has bachelor's or master's degree but neither parent has higher, and at least one parent has Ph.D. or equivalent.

TABLE 10

X3TGPATOT	numbers	percent
4	797	0.049
3.5	2467	0.152
3	3186	0.197
2.5	3202	0.198
2	2554	0.158
1.5	1488	0.092
1	806	0.050
0.5	354	0.022
0.25	143	0.009
-8	1164	0.072
-9	28	0.002

GPA distribution of students having parents without a bachelor's degree or higher equivalent.

TABLE 11

X3TGPATOT	numbers	percent
4	961	0.155
3.5	1903	0.306
3	1446	0.233
2.5	889	0.143
2	408	0.066
1.5	154	0.025
1	57	0.009
0.5	24	0.004
0.25	4	0.001
-8	355	0.057
-9	17	0.003

GPA distribution of students having parents with at least a bachelor's but no more than a one master's degree or equivalent.

TABLE 12

X3TGPATOT	numbers	percent
4	287	0.262
3.5	363	0.331
3	225	0.205
2.5	93	0.085
2	41	0.037
1.5	15	0.014
1	5	0.005
0.5	3	0.003
0.25	1	0.001
-8	56	0.051
-9	7	0.006

GPA distribution of students having at least one parent with a doctorate degree or equivalent.

The parent's educational level heavily affects the student's performance. In the first category, where both the student's parents did not have a college education, the student's GPAs are spread around relatively evenly. In the second category where at least one parent has a bachelors or masters degree, the percentages are concentrated around the top of the GPA scale, and the same is true to even more of a degree with the last category where at least one of the student's parents has a Ph.D. or equivalent.

5. How AP courses taken relate to high school student's GPA?

In the original data set, credits earned in AP courses range from 0 to 13. We calculate the average GPA for each credit level. Following shows the SQL query and the corresponding output:

```
SELECT X3TCREDAPIB AS AP_credit, ROUND(AVG(X3TGPATOT),3) AS avg_GPA
FROM export_dataframe
GROUP BY X3TCREDAPIB
HAVING COUNT(*)>1
ORDER BY X3TCREDAPIB;
```

Fig 4. MySQL code for querying GPA average with respect to AP/IB credits

TABLE 13

AP_credit	avg_GPA		
0	2.372	7	3.555
1	3.043	8	3.594
2	3.248	9	3.609
3	3.338	10	3.601
4	3.381	11	3.583
5	3.482	12	3.602
6	3.519	13	3.588

Average GPA of students taking various numbers of AP credits.

Result: Average GPA for students increases monotonously as credits earned in AP courses range from 0 to 9. Average GPA remains at a high level after AP courses credits reach 9.

IV. CONCLUSION

a) Summary

In this project, we have completed a series of queries related to factors that affect high school students' academic performance and given some suggestions based on the results. We used the High School Longitudinal Study data set, which has entities such as students, parents and schools, and their related features. To accommodate our research interest, we preprocessed the data set with pandas and selected variables of interest. To complete these queries, we used MySQL workbench and constructed queries depending on different data types of variables of interest. We also have post-process steps for some queries to do error-handling.

b) Limitation

Currently, we only consider variables that we intuitively think are related to a student's academic performance. In order to obtain more rigorous and comprehensive results, we can first do correlation analysis for all variables and find out variables that correlate strongly with students' academic performance. In addition, to further improve the performance of querying, we can add indexes to the frequently queried fields. In this way, we do not need a full table scan and can realize more rapid lookups.

V. REFERENCES

[1] Pew Research Center (2017). "U.S. students' academic achievement still lags that of their peers in many other countries", retrieved from https://www.pewresearch.org/fact-tank/2017/02/15/u-s-students-internationally-math-science/

[2] Politifact, The Poynter Institute. "Back to school: Is the United States falling behind on education?", retrieved from https://www.politifact.com/article/2018/aug/20/back-school-united-states-falling-behind-education/

[3] High School Longitudinal Study of 2009. Retrieved December 14, 2020, from https://catalog.data.gov/dataset/high-school-longitudinal-study-of-2009

[4] HSLS:09. Retrieved December 14, 2020, from https://nces.ed.gov/surveys/hsls09/