DATA ANALYST 3 - ADVANCED SQL

PROJECT PORTFOLIO REPORT

BY HENRY UGOCHUKWU OKAM

Table of Contents

1. Introduction/Project background 3
2. Root cause analysis process.. 4
3. Insights from the dataset.. 5
    a. SQL Codes Utilized.. 5
    b. Key Insights... 8
4. Visualizations of these insights.. 10
    a. Reporting Dashboard... 17
5. Important findings and recommendations to solve the business problem.................. 18
6. Conclusion...19

Introduction to the dataset

The objective of this project is to evaluate the factors affecting students educational performance.

In order to realize this objective, four csv files were provided by *Entry Level* for the purpose of this report. The four csv files contain data about the country info, student academic info, student family details and student personal details respectively for 4739 unique student id. These four files which are related contains key columns names such as :

gender: factor indicating gender.

ethnicity: factor indicating ethnicity (African-American, Hispanic, Asian or other).

academic_score: student's academic score throughout high school and college

student_tuition: cost of tuition for the student

education: the years of education the student has received

fcollege: factor. Is the father a college graduate?

mcollege: factor. Is the mother a college graduate?

home: factor. Does the family own their home?

urban: factor. Is the school in an urban area?

unemp: county unemployment rate in 2020

income: high or low income household based on county average

wage: state hourly wage in manufacturing in 1980

distance: distance from 4-year college (in 10 miles)

region: factor indicating region (West, East or other)

avg_county_tuition: average state 4-year college tuition (in 1000 USD)

After data collection, all four csv files were imported into *sqlite online* and combined to form one reporting table known as *reports_student_colleges* with a total of 37,997 student records. It was observed that there is a one to many relationship between the *student_personal_details* table which contained unique student records and the other three tables when combined.

From the *reports_student_colleges* table, several tables were further created using SQL queries for the purpose of answering our research questions. These tables were then exported to google sheets, further wrangled , then used to create compelling visualizations to give insights to the problem statement.

Root Cause Analysis Process

In order to ascertain the varying factors that may have an effect on a student's academic performance , the following questions were asked of our data.

1. What are the proportion of educated students by ethnicity
2. How can we evaluate the Cost of college tuition against household income
3. What is the effect of a student's parents education on the education of the student
4. Is location of school a determinant on tuition
5. What is the differences in student's performance with respect to where the school is location
6. Which ethnic group has a higher performance rate in schools and which gender performs better on average?
7. Does the age of a student have an effect on their academic performance in school?

In order to answer these questions and for the purpose of this analysis, some assumptions were made. One is that the student data collected represents sample data for high schools/colleges in the United State of America (USA). This is due to the diversity and multicultural nature of the USA student community. We also assumed that academic scores greater than or equal to 50% would be deemed as a 'pass' while academic scores less than 50% would be deemed as a 'fail' for the purpose of this study. And lastly we assumed that the ***education*** variable indicates the years of education received by the student since birth.

Insights from the dataset

a) SQL Codes Utilized

TABLE 1

Combining all tables to create reporting table called reports_student_colleges

```
CREATE TABLE reports_student_colleges AS

SELECT student_personal_details.id AS id,

      , gender

      , DOB

      , ethnicity

      , academic_score

      , student_tuition

      , education

      ,unemp

      ,wage

      ,distance

      ,region

      ,avg_county_tuition

      ,fcollege

      ,mcollege

      ,home

      ,urban

      ,income

FROM student_personal_details

  LEFT JOIN student_academic_info

    ON student_personal_details.academic_info_id = student_academic_info.id

  LEFT JOIN county_info
```

```
    ON student_personal_details.county_id = county_info.id
  LEFT JOIN student_family_details
    ON student_personal_details.family_details_id = student_family_details.id
ORDER BY 1;
```

TABLE 2

What is the proportion of educated students by ethnicity ?

```
CREATE TABLE education_by_ethnicity AS
SELECT COUNT(id) AS no_of_student
        , ethnicity
        , AVG(education)  AS avg_education
        , AVG(academic_score)AS avg_academic_score
FROM reports_student_colleges
GROUP BY 2
```

TABLE 3

How can we evaluate the Cost of college tuition against household income ?

```
CREATE TABLE tuition_by_income AS
SELECT count(*) AS student_count
        , AVG(student_tuition) AS avg_student_tuition
       , AVG(avg_county_tuition) as avg_county_tuition
       , income
FROM reports_student_colleges
GROUP BY 4
ORDER BY 1 DESC;
```

TABLE 4

What is the effect of a student's parents' education on the education of the student ?

```
CREATE TABLE education_status AS
SELECT  COUNT (DISTINCT ID) as student_id_count
             , AVG(academic_score) AS avg_score
             , AVG(education) AS avg_education
             , fcollege
            , mcollege
FROM  reports_student_colleges
GROUP BY 4,5
ORDER BY 1,2,3;
```

<u>TABLE 5</u>

-Is location of school a determinant on tuition ?

-What are the differences in a student's performance with respect to where the school is located ?

```
CREATE TABLE school_location_report AS
SELECT AVG(academic_score) AS academic_score
            , region
            , urban
            , AVG(student_tuition) AS student_tuition
            , AVG(avg_county_tuition) AS avg_county_tuition
FROM  reports_student_colleges
GROUP BY 2,3
```

<u>TABLE 6</u>

Which ethnic group has a higher performance rate in schools and which gender performs better on average?

```
CREATE TABLE ethnic_gender_report AS
SELECT  ethnicity
            , gender
            , AVG(academic_score)  AS academic_score_by_ethnicity
FROM reports_student_colleges
GROUP BY 1,2;
```

<u>TABLE 7</u>

Does the age of a student have an effect on their academic performance in school?

```
CREATE TABLE students_status_by_age
SELECT id
           , dob
           , strftime('%Y','now') AS current_year
          , academic_score
          , CASE WHEN academic_score >= 50 then 'Pass'
                    ELSE 'Fail' END AS grade
FROM reports_student_colleges;
```

b) Key Insights

- The African American students are the least represented while the 'other' category of students are the most represented in our student's sample.

- Average number of years spent by students in school is 14 years for students of all ethnicities.

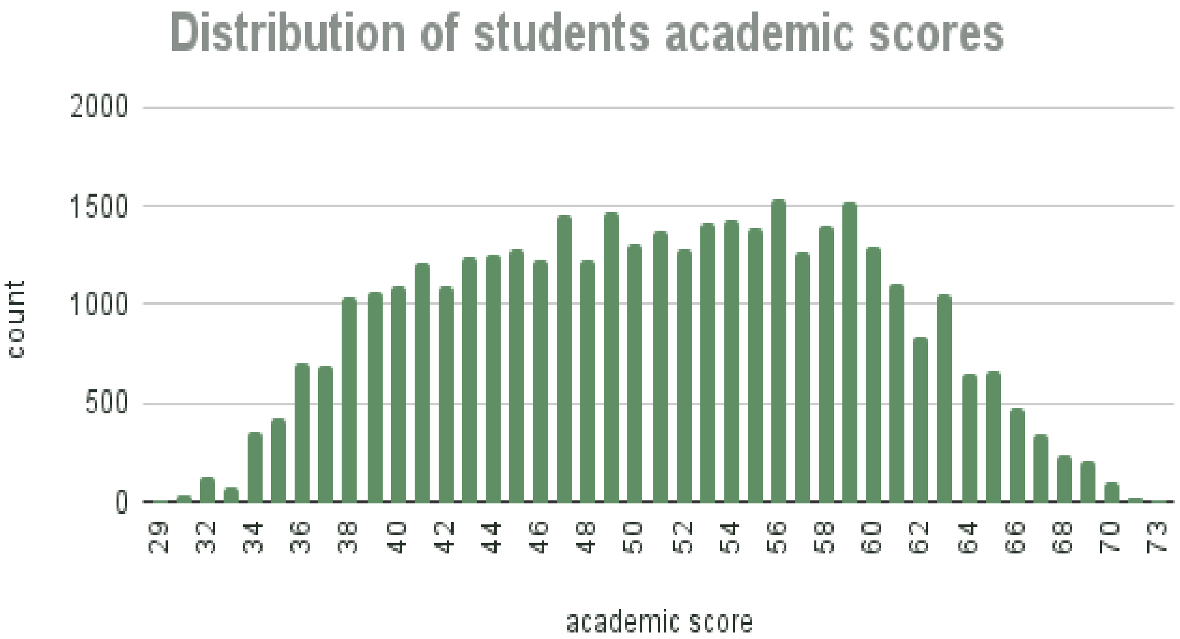
- There was no considerable effect of a student's age on their academic scores.

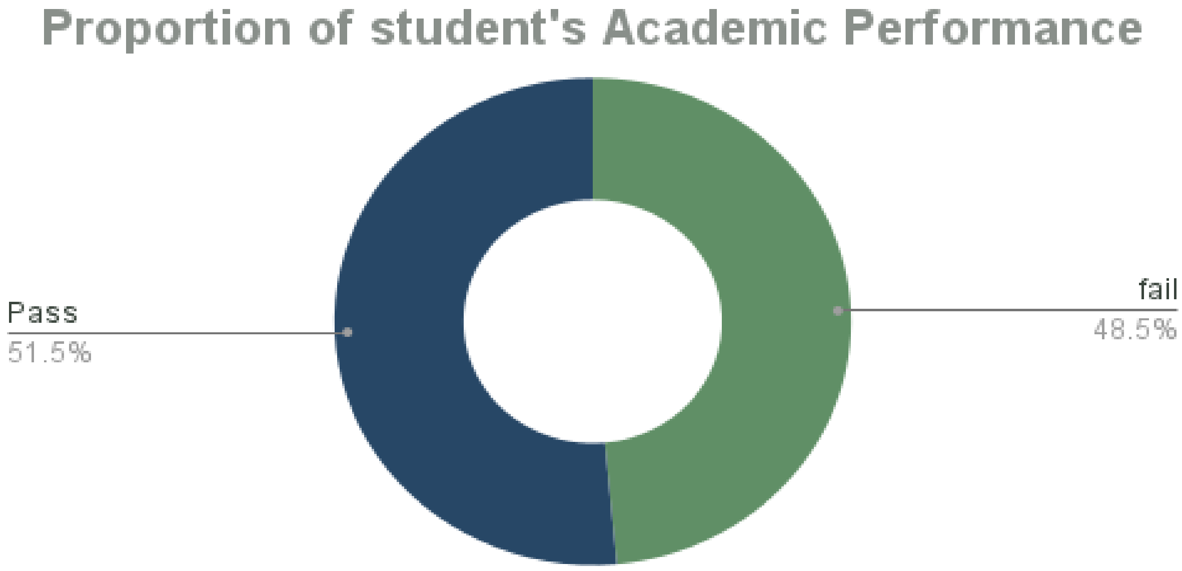
- Students from high income homes pay on average slightly higher student tuition (0.72% higher) than those from low income homes.

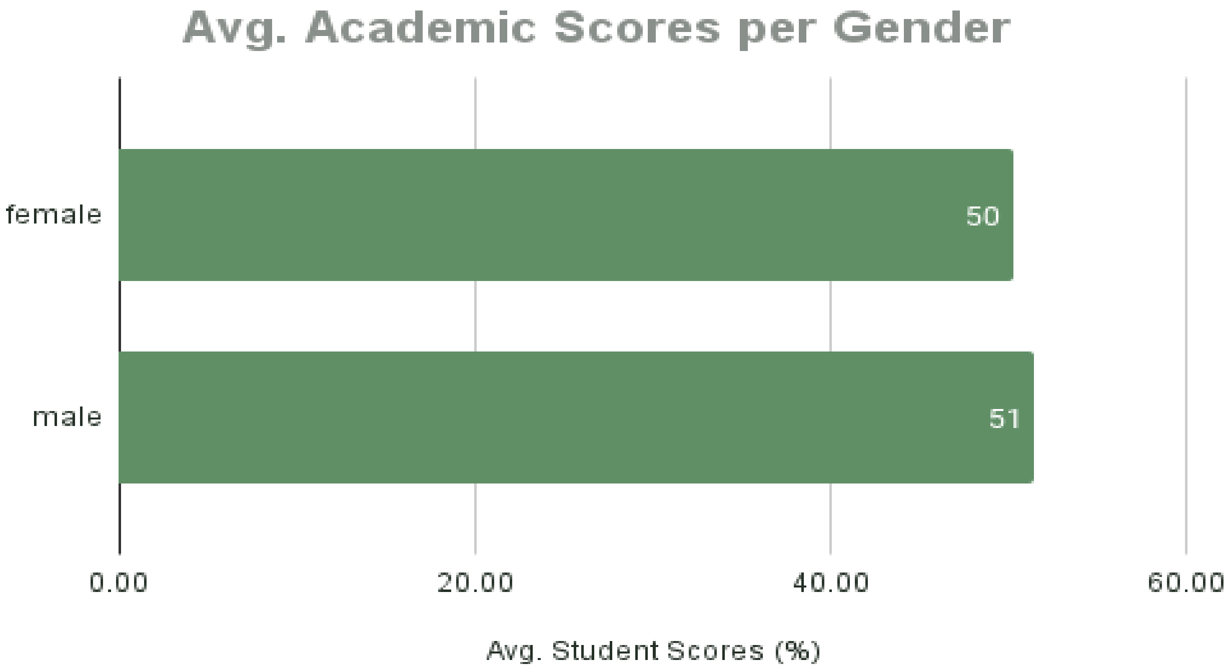
- Students from high incomes homes pay a country tuition of averagely 1.23% higher than those from low-income homes

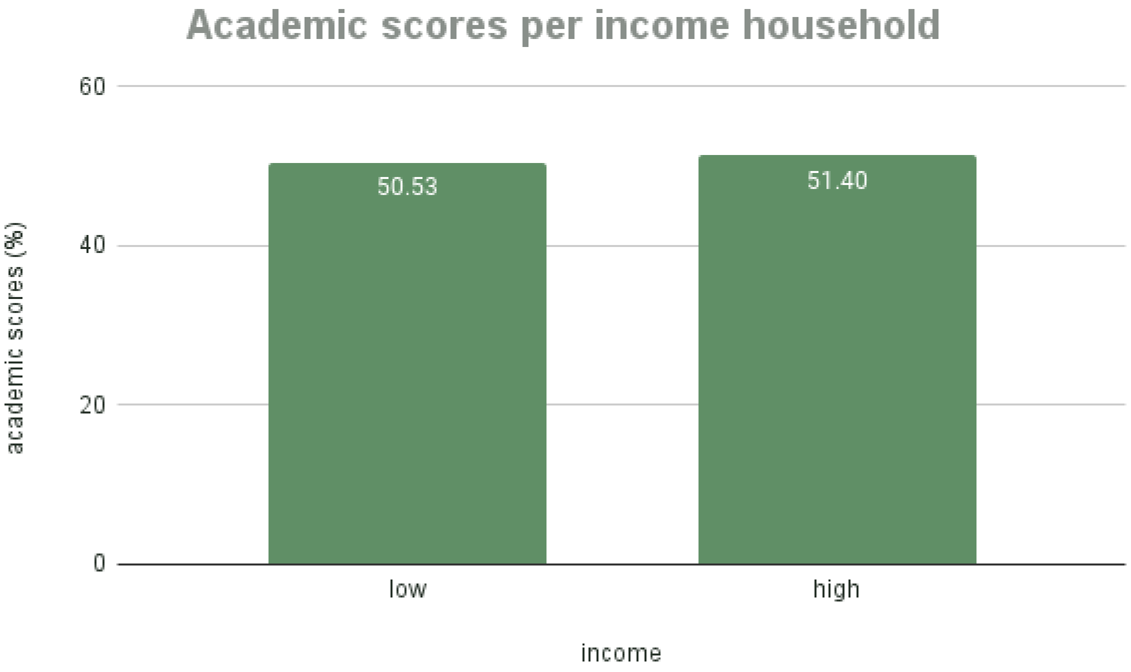
- There is no considerable effect of parents education on the education of the student
- Student who have both parents educated perform better in academic scores and have more years of education than those with both or either parent uneducated
- Students with both parents not educated perform the least in academic scores and have lesser years of education than those with both or either parents educated
- Average country tuition for the western region of the country is very much lower than other regions of the country.
- There is no considerable difference in academic scores of students in the different regions/settlement type
- Average academic scores per student across all region is 51%
- Male students on average have better academic scores than female students
- In terms of academic performance, Asians rank better than all other ethnic groups in the country.
- A plot of academic scores for all students shows a symmetrical distribution
- More students have an academic score of 56% than any other score.
- Lowest academic score of any student is 29% and highest of any student is 73%.
- 51.5% of all students passed (scored greater than or equal to 50) whilst 48.5% of all students failed (scored less than 50).

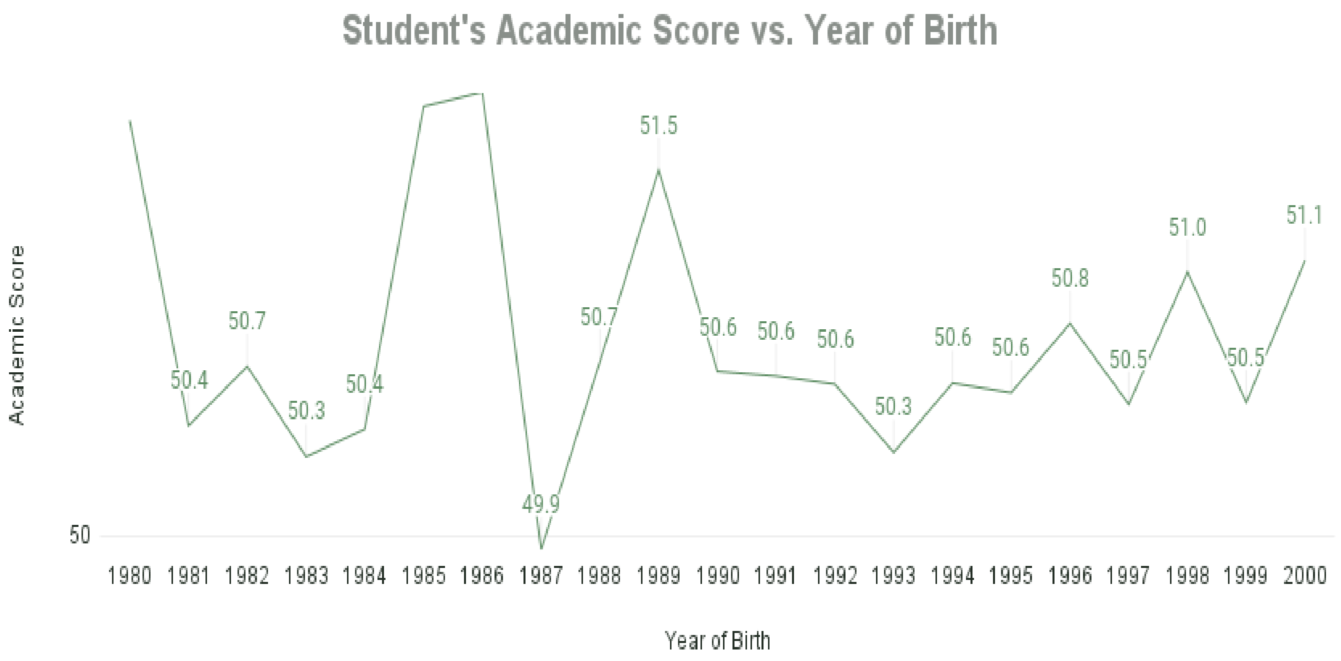
Visualizations

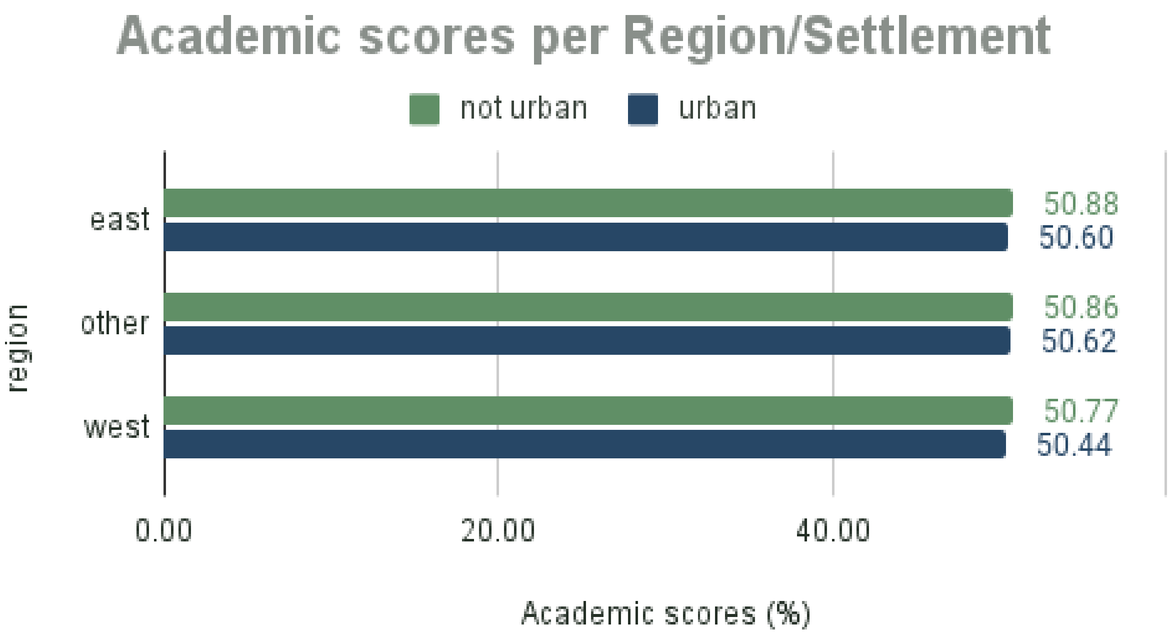


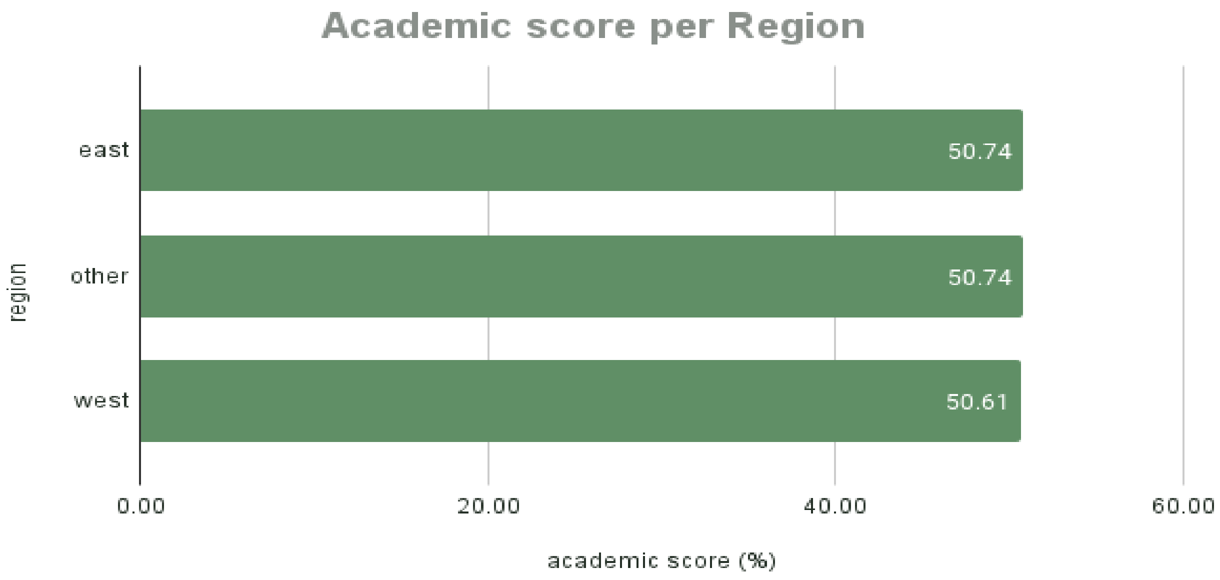


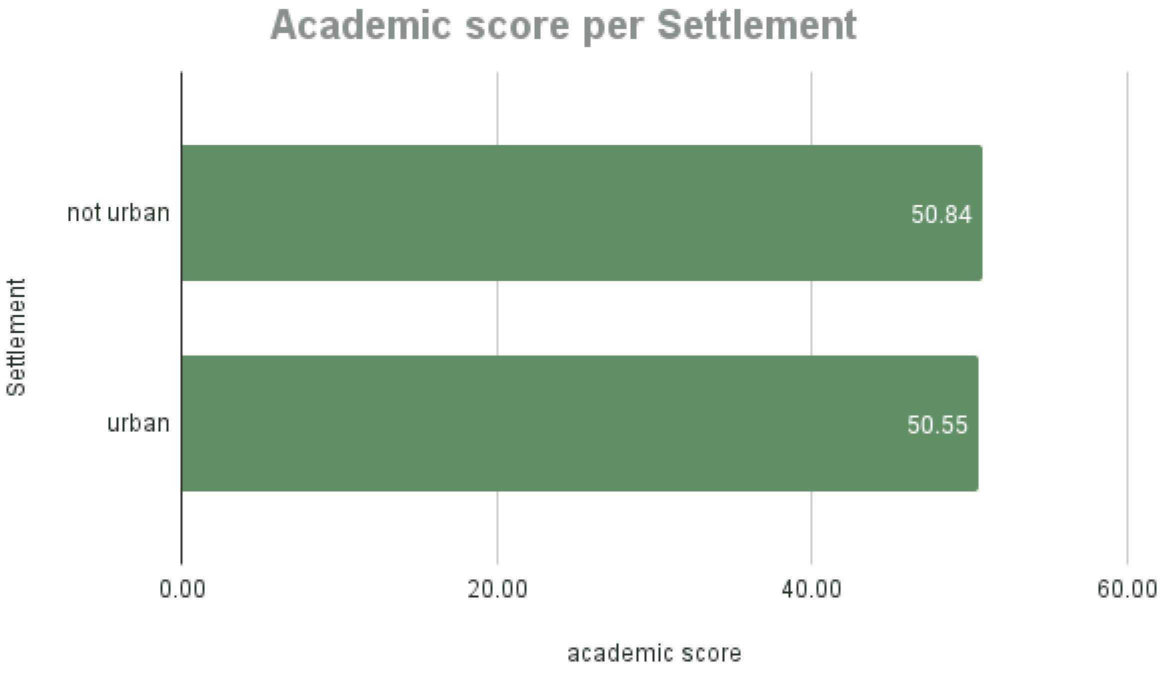

Avg. Academic Scores per Gender
female
50
male
51
0.00
20.00
40.00
60.00
Avg. Student Scores (%)

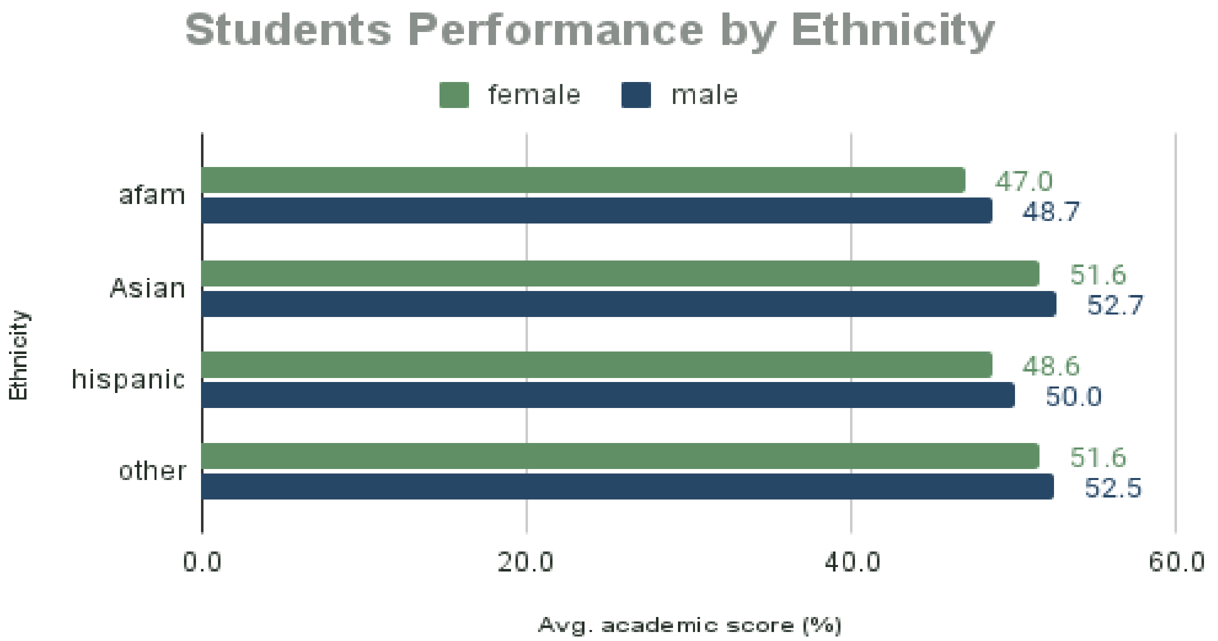

Academic scores per income household
60
40
20
0
academic scores (%)
50.53
51.40
low
high
income

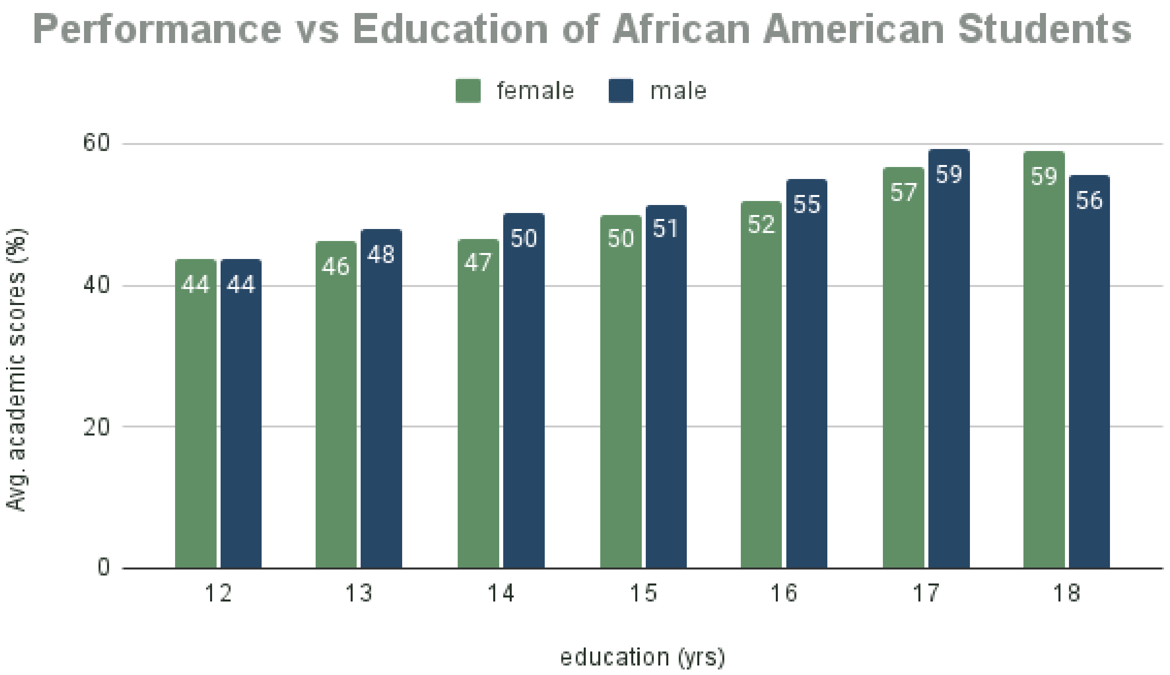

Student's Academic Score vs. Year of Birth
Academic Score
50
50.4
50.7
50.3
50.4
49.9
50.7
51.5
50.6
50.6
50.6
50.3
50.6
50.6
50.8
50.5
51.0
50.5
51.1
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
Year of Birth

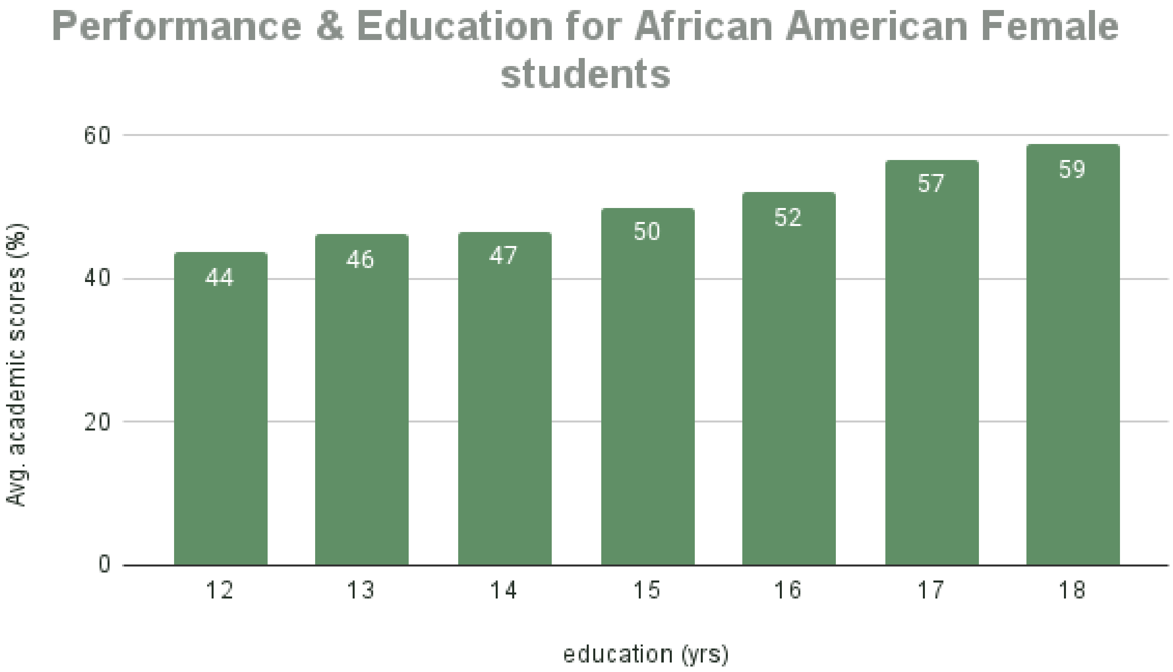

Academic scores per Region/Settlement
not urban
urban
region
east
50.88
50.60
other
50.86
50.62
west
50.77
50.44
0.00
20.00
40.00
Academic scores (%)


Academic score per Region
region
east
other
west
50.74
50.74
50.61
0.00
20.00
40.00
60.00
academic score (%)

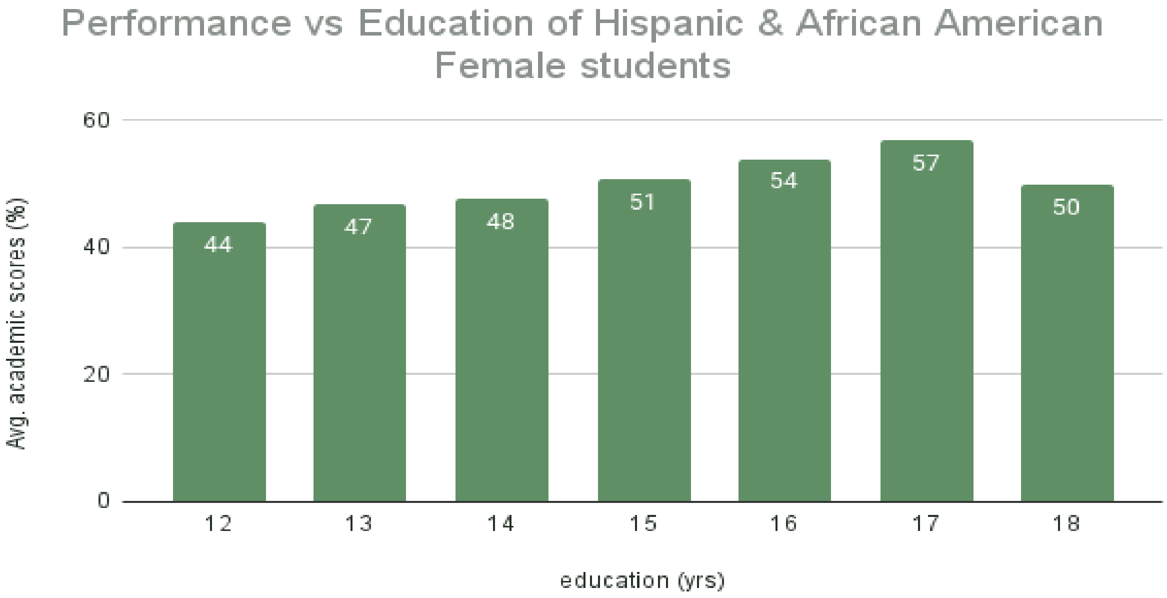

Academic score per Settlement
Settlement
not urban
urban
50.84
50.55
0.00
20.00
40.00
60.00
academic score


Students Performance by Ethnicity
female
male
Ethnicity
afam
47.0
48.7
Asian
51.6
52.7
hispanic
48.6
50.0
other
51.6
52.5
0.0
20.0
40.0
60.0
Avg. academic score (%)

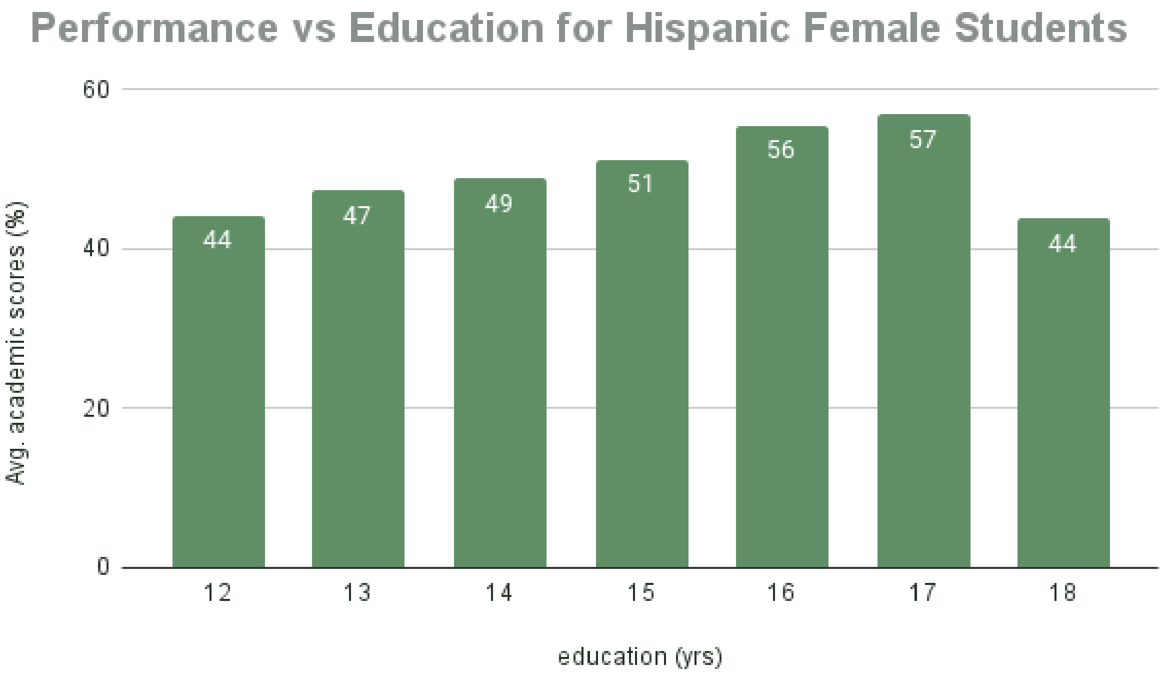

Performance vs Education of African American Students
female
male
Avg. academic scores (%)
60
40
20
0
44
44
46
48
47
50
50
51
52
55
57
59
59
56
12
13
14
15
16
17
18
education (yrs)

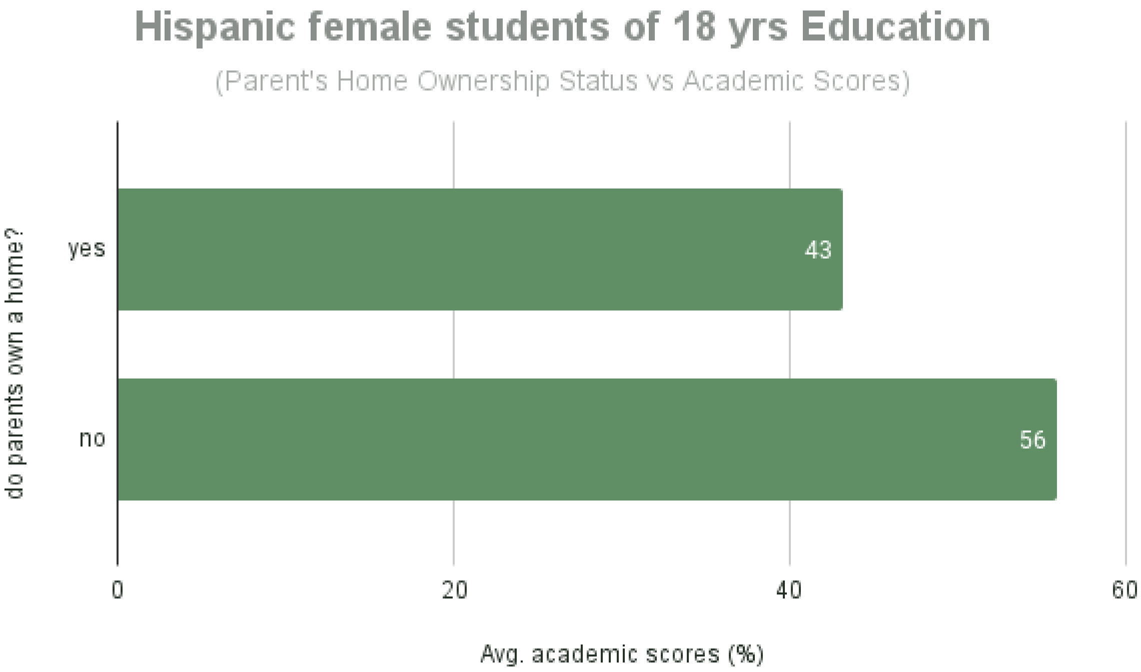


Performance & Education for African American Female students

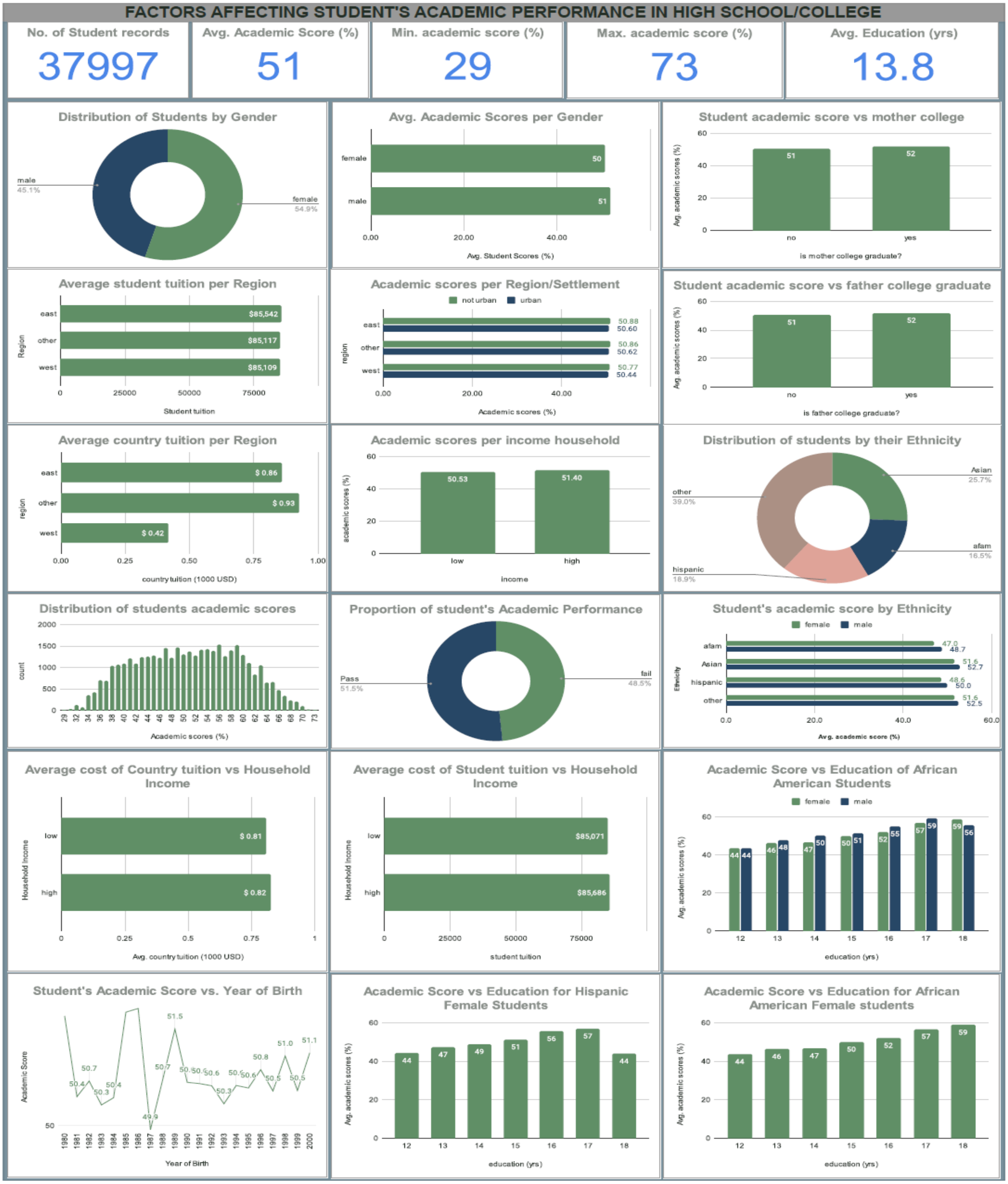


Performance vs Education of Hispanic & African American Female students


Performance vs Education for Hispanic Female Students
Avg. academic scores (%)
60
40
20
0
44
47
49
51
56
57
44
12
13
14
15
16
17
18
education (yrs)


Hispanic female students of 18 yrs Education
(Parent's Home Ownership Status vs Academic Scores)
do parents own a home?
yes
43
no
56
0
20
40
60
Avg. academic scores (%)

<u>**Reporting Dashboard**</u>



Important findings/Recommendations

From the problem statement which is to evaluate the factors affecting the student's academic performance in high school/colleges in the USA , the visuals clearly shows that the overall average academic score for students is 51%. This is a rather low average score which begs the question of why we have a low average student academic score.

Also from the visuals , it can be ascertained that the average academic score for students with an ethnicity of african american is 47.85% which shows a high failure rate for these categories of students. In terms of gender, male African American students have an average score of 48.7% while females have an average score of 47%, both of which show a high failure rate amongst these categories of students.

Also, hispanic female students share in this high failure rate as seen from their average academic score of 48.6%. Thus we can therefore infer that the average academic scores of the students was of a low rate due to the high failure rate of African American students and hispanic female students.

If we then proceed to find out why the academic scores of african american students and hispanic female students are poor, we would observe that only the academic scores of female african american students are poor for students with an overall education spanning 12, 13, and 14 years respectively. Male african american students share in this poor standing for students with an education of 12 and 13 years respectively.

The same pattern can also be observed for female hispanic students with an education period of 12, 13, 14 and 18 years.

We can however infer from the visuals plotted that the more years of education acquired by a student, the greater their academic scores. An exception to this insight would be the hispanic female students with 18 years of education plus a poor academic score. This insight when closely inspected, showcases that the high failure rate is more concentrated on those female hispanic 18 years of education students whose parents own a home. Other than that observation, there is no considerable reason from the data why female hispanic students with 18 years of education would record such a high failure rate.

Based on these aforementioned insights as garnered and the corresponding analysis, it is recommended that school authorities and its tutors pay better attention to its foreign student's academic performance and other non-academic factors such as racism, mental health, emotional health, adaptability, guidance & counseling and other aspects of their foreign student's lives that may have an adverse effect on their academics when not properly managed. This is even more necessary especially for female students from african american and hispanic ethnicities.

Conclusion

The academic performance of high school/college students in the USA was analyzed and properly assessed from the available student's record data based on a number of factors such as the academic score attained all through a student's high school/college, the number of years of education the student has received since birth, the student's age, student's tuition, average country tuition, location of school, ethnicity of student, student's household income, gender, and student parent's college education status.

An analysis on all these aforementioned criteria brought about a couple of key insights which formed the basis of this report.

Based on the insights garnered, it was ascertained that from a student record of 37, 997 student data analyzed, only 51.5% of students have a pass score whilst 48.5% of students have a failed score with an average academic score of 51%. Further analysis to ascertain the reason for a relatively high rate of low scores amongst students revealed that most fail scores came from students of hispanic and african american ethnicities while students from an asian ethnicity have a higher pass rate on average. It was observed that the female gender amongst both hispanic/african american students occupied a larger percentage of the failure rate. It was also observed that an increase in the years of education received by a student has a corresponding effect on the student's academic score with the exception of female hispanic students with 18 years of education.

However, In the course of analyzing the student data, certain limitations were encountered. One of such limitation was insufficient information from the dataset to properly ascertain which category of student records were made available, and this led to a couple of assumptions including - *the country where the data was gathered*, *the pass mark for the student's academic score, and the time period indicated by the* ***education*** *variable in the data.*

Additional information on these parameters would have come in handy to better understand the dataset. Also, some factors which could affect a student's academic performance were not included in the data. These factors which may be also useful in ascertaining either the success rate or failure rate of a student includes but are not limited to - *health status of student, information on high school/colleges attended, university course of study (where applicable), student status(international/local), country of residence* etc.

More clarity on the cause of success or failure could be better ascertained if these additional information were provided.