Colin Michael

DSC 680: Project 2 Milestone 2

April 24th, 2024

School Attendance by Student Groups

Business Problem:

Public education is the one of the most important parts of society. It is crucial that young people are afforded the opportunity to learn and prepare themselves for successful lives. Students need to have consistent attendance to get the most value out of school. My research paper will focus on applying data science towards improving student education in public schools.

Background/History:

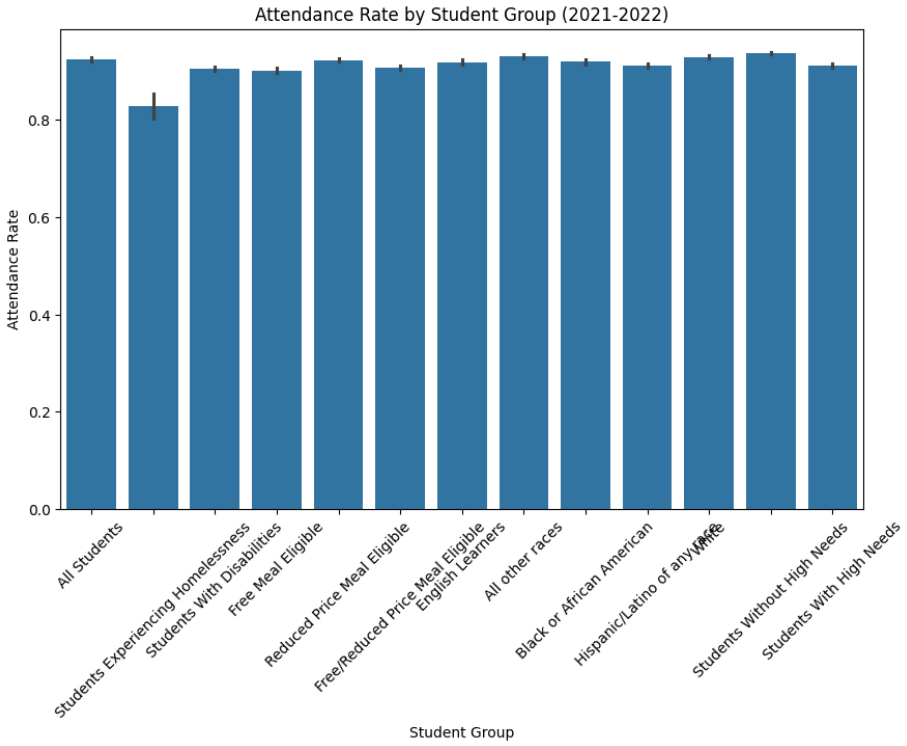
My second project in DSC 680 will be focused on school attendance rate by students in Connecticut public schools' grades PK-12 for the 2020-2021 and 2021-2022 school years. I will be assuming the role of a data scientist hired by the State of Connecticut. My primary goal is to utilize the existing data on school attendance to provide recommendations to Connecticut on how to improve school attendance. I want to draw actionable insights from the data to help Connecticut identify at-risk segments of students.

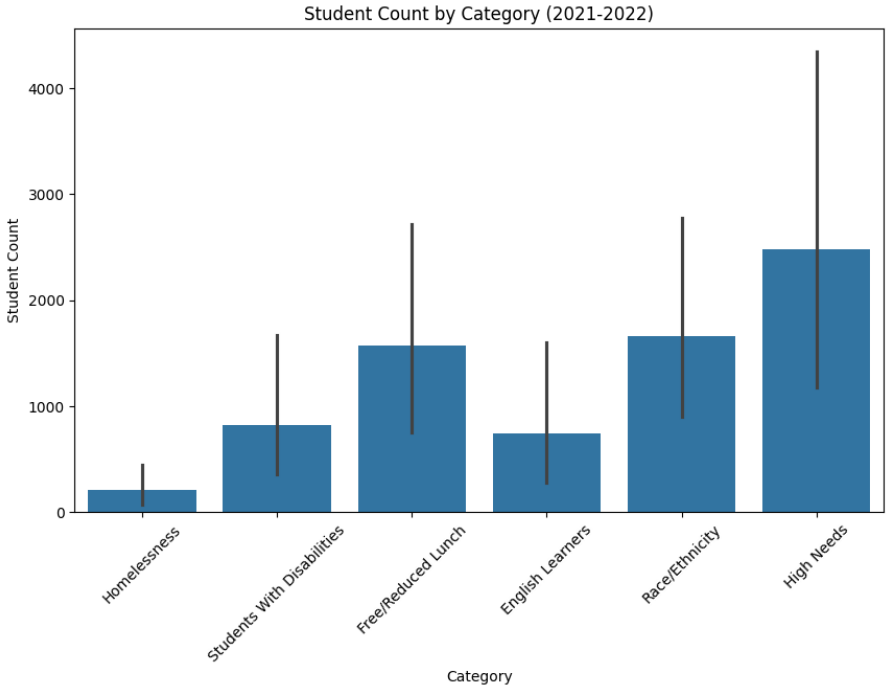
Data Explanation:

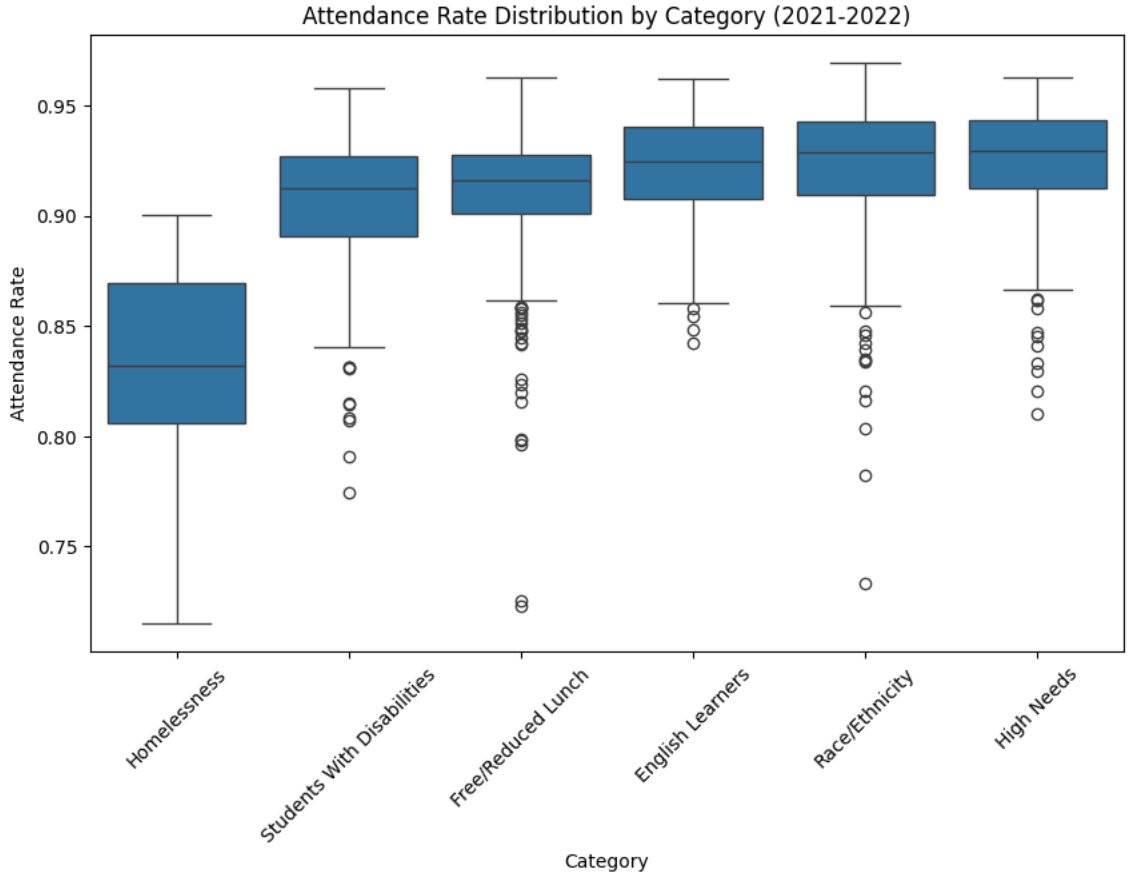
I will be analyzing the impact that different demographic factors have on school attendance. My primary dataset is a CSV file from Kaggle.com titled 'Student School Attendance' (Kaggle.com). It includes 2,000 rows of data, with each row of data providing enrollment and attendance numbers for different segments of students across all the school districts in Connecticut. Each row falls into a category of homelessness, students with disabilities, free/reduced lunch, English learners, race/ethnicity, and high needs. Sub-categories

include all students, students experiencing homelessness, students with disabilities, free meal eligible, reduced-price meal eligible, free/reduced price meal eligible, English learners, Black or African American, Hispanic/Latino, White, All other races, students with high needs, and students without high needs.

Data Visualizations:




Student Count by Category (2021-2022)
Student Count
0
1000
2000
3000
4000
Homelessness
Students With Disabilities
Free/Reduced Lunch
English Learners
Race/Ethnicity
High Needs
Category


Attendance Rate Distribution by Category (2021-2022)
Attendance Rate
0.75
0.80
0.85
0.90
0.95
Homelessness
Students With Disabilities
Free/Reduced Lunch
English Learners
Race/Ethnicity
High Needs
Category

Methods, Analysis, and Conclusions:

First, I will apply a regression analysis to the data. Predicting future attendance rates is the primary aim of my regression. I performed an Ordinary Least Square regression on my dataset in Python. I used 2021-2022 attendance rate as my dependent variable because that is what I am trying to evaluate. My independent variables were 2020-2021 student count, 2021-2022 student count, and all the student groups. These variables are categorical, so I had to convert them to dummy variables.

Here are my OLS regression results. My model produced an R-squared of .257, which reflects that 25.7% of the variation in attendance rates are explained by the model. The p-value is used to show statistical significance of the variables. We are ideally looking for a p-value under .05. The following variables show a negative relationship towards 2021-2022 attendance with a significant p-value: Black or African American, English Learners, Free Meal Eligible, Free/Reduced Priced Meal Eligible, Hispanic/Latino, Reduced Meal Price Eligible, Students Experiencing Homelessness, Students with Disabilities, White. Students Experiencing Homelessness, Free Meal Eligible and Students with Disabilities had the highest negative relationship with attendance.

```
                                 OLS Regression Results
==============================================================================================
Dep. Variable:     2021-2022 attendance rate - year to date   R-squared:                       0.257
Model:                                                 OLS   Adj. R-squared:                  0.251
Method:                                      Least Squares   F-statistic:                     44.49
Date:                                     Wed, 24 Apr 2024   Prob (F-statistic):          1.52e-105
Time:                                             22:21:14   Log-Likelihood:                 4084.8
No. Observations:                                     1818   AIC:                            -8140.
Df Residuals:                                         1803   BIC:                            -8057.
Df Model:                                               14
Covariance Type:                                 nonrobust
==============================================================================================
                                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                                        0.8993      0.004    219.309      0.000       0.891       0.907
2021-2022 student count - year to date -4.039e-08   1.56e-06     -0.026      0.979    -3.1e-06    3.02e-06
2020-2021 student count                -7.301e-08   1.55e-06     -0.047      0.962   -3.11e-06    2.96e-06
2020-2021 attendance rate                    0.0361      0.004      9.644      0.000       0.029       0.043
Black or African American                   -0.0115      0.003     -3.541      0.000      -0.018      -0.005
English Learners                            -0.0118      0.003     -3.549      0.000      -0.018      -0.005
Free Meal Eligible                          -0.0284      0.003     -9.122      0.000      -0.034      -0.022
Free/Reduced Price Meal Eligible            -0.0231      0.003     -7.787      0.000      -0.029      -0.017
Hispanic/Latino of any race                 -0.0191      0.003     -6.230      0.000      -0.025      -0.013
Reduced Price Meal Eligible                 -0.0089      0.003     -2.815      0.005      -0.015      -0.003
Students Experiencing Homelessness          -0.0952      0.007    -13.459      0.000      -0.109      -0.081
Students With Disabilities                  -0.0260      0.003     -8.714      0.000      -0.032      -0.020
Students With High Needs                    -0.0197      0.003     -6.628      0.000      -0.026      -0.014
Students Without High Needs                  0.0040      0.003      1.332      0.183      -0.002       0.010
White                                       -0.0030      0.003     -0.992      0.322      -0.009       0.003
==============================================================================
Omnibus:                      503.758   Durbin-Watson:                   0.590
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2099.199
Skew:                          -1.279   Prob(JB):                         0.00
Kurtosis:                       7.601   Cond. No.                     2.63e+05
==============================================================================
```

Assumptions:

I am making several assumptions in my data regression analysis. First, I am assuming a linear relationship between 2020-2021 attendance and the independent variables. I am also assuming that my variables are not multicollinear. For example, I am assuming that students' status with or without high needs is not correlated to free meal eligibility.

Limitations:

My dataset only includes attendance data for 2020-2022. My paper would be stronger if I had more data able to be used. Also, the COVID-19 pandemic lockdown took place during my dataset. Lockdowns dramatically impacted school attendance, with various measures aimed at reducing the spread of the virus. I am worried the pandemic will skew my results.

Challenges:

One of the biggest challenges to my paper is that the dataset does not include economic factors on the school regions. I would be curious to add in economic conditions to my regression analysis to see if things like average household income impacts school attendance.

Future Uses/Applications:

I would be excited to apply this dataset to other states to see what their school attendance patterns look like. I would like to evaluate if different regions of the United States have different variable impact on K-12 attendance. It would also be interesting to dive into specific age/grades of the dataset to see how different dependent variables impact different aged students.

Recommendations:

Based on my linear regression, I would advise the State of Connecticut to allocate additional resources to at-risk groups of the student population. Specifically, students on free or reduced lunch plans should receive additional support in getting to school.

Implementation Plan:

My next steps are looking into the dataset to see if I can improve or run an additional supporting regression. I would like to evaluate more variables and try to reach a higher statistical significance level.

Ethical Considerations:

I am concerned that the findings from my paper could lead to negative prejudices towards at-risk groups. It is crucial that my findings are presented in a way that supports, rather than knocks down, different groups of society. Education is a basic human right, and I am dedicate towards making sure my paper is used to help people.

10 Questions:

1. Why did you choose this project?: Education has had such a massive impact on me, and I want to do my part in furthering the educational system in America.
2. What surprised you the most about your findings?: The impact of disabilities on attendance
3. What additional data point would you be interested in?: Age of students to see how grade and age impacts attendance
4. What is the biggest limitation of the paper?: Only using data from Connecticut
5. What other States in the U.S. would you want to compare Connecticut to?: A state from each region of the U.S.
6. What are other data science methods you could apply to the project?: A time series forecast
7. What is the single biggest takeaway from the study?: The massive impact income-related variables have on school attendance
8. How would you improve the project to account for COVID-19?: Pull in more years of data
9. Do you think private schools would have similar trends?: I think the income barriers to private schools would massively change the results.
10. What is the impact on school size and attendance?: I would want to pull in town population

Appendix

There are many more resources available on the topic of school attendance for youths. A French-organization conducted a study on international school attendance. The U.S. overall has a

lower absentee rate than Canada. Japan has the lowest in the world, sitting at 4% (TheGlobeAndMail.com)

The United States Department of Education conducted an analysis on absenteeism across the United State. Connecticut, which is where my study focuses on, has one of the lowest absenteeism rates. All New England has low absenteeism rates. Alaska has the highest absenteeism rates in the nation. (Ed.Gov)

The COVID-19 pandemic caused widespread lockdowns in the United States. Many schools added remote options to lower the spread of the virus. The New York Times published an article in April 2024 digging deeper into the pandemic's lasting impact on absenteeism. Duke University Professor Kate Rosanbalm argued that student's relationship with schools fundamentally changed during that period, and attitudes toward attendance have not recovered (NYTimes.com).

Citations

Alphonso, C. (2003, October 15). *Canada lags in school-attendance test*. The Globe and Mail. https://www.theglobeandmail.com/news/national/canada-lags-in-school-attendance-test/article18432952/

Chronic absenteeism in the nation's schools. Chronic Absenteeism in the Nation's Schools. (n.d.). https://www2.ed.gov/datastory/chronicabsenteeism.html#three

Network, T. L. (2024, April 11). *What students are saying about why school absences have "exploded."* The New York Times. https://www.nytimes.com/2024/04/11/learning/what-students-are-saying-about-why-school-absences-have-exploded.html