Case Study 1:
Insights from Early Childhood Education Data in the School District of Beloit[1]

Report Compiled By:

Aaditya Joshi' 25
Quantitative Economics & Biochemistry Major, Minor in Finance

Prince Upadhyay' 25
Data Science

Richard J Pouzar' 25
Data Science

[1] OpenAI. (2024). *ChatGPT* (November version) [Large language model]. https://chatgpt.com/share/67317dc6-d4e4-8005-9f7b-ac20eb239778.

ABSTRACT[2]

This study examines the impact of various demographic and educational factors on the academic performance of third-grade students within the Beloit School District, which currently ranks among the lowest-performing districts in terms of academic achievement. Utilizing data gathered from 213 students' parents surveyed during a parents' conference hosted across the district's ten schools, we analyze the effects of school designation, special education status, English language proficiency, and duration of residency in Beloit on academic outcomes. Academic performance is assessed through the Measures of Academic Progress (MAP)[3] examinations in Mathematics and Reading. A linear regression model was employed to evaluate the contribution of each factor to student performance, thereby offering results into areas requiring attention to enhance educational outcomes. The findings highlight the intricate interplay of socio-educational variables in influencing student success and provide a foundation for targeted interventions aimed at improving academic achievement in underperforming school districts. Policies and recommendations are suggested based on the results of the regression model.

I. INTRODUCTION

The Beloit School District has consistently received low rankings in academic performance within Wisconsin, a situation correlated with its designation as one of the state's most economically disadvantaged districts. In 2019, Professor Diep Phan, in collaboration with student research assistants, conducted a comprehensive survey aimed at evaluating parental demographics and their involvement in the educational processes of their children. Professor Phan disseminated the survey findings, along with associated test scores and demographic data of the participating student population, facilitating an investigation into potential contributors to suboptimal test scores. Initial analyses indicated several pertinent factors: the exclusive use of an English-language testing format may disproportionately affect students with diverse linguistic backgrounds, and variables such as the duration of family residency in Beloit and attendance at elementary schools could significantly impact educational outcomes. We applied linear

[2]We recognize that the inspiration and selection of this topic were influenced by the DSDA 300: Senior Seminar course, under the guidance of Dr. Phan and Dr. Disha. We also extend our gratitude to Dr.Phan for providing us with the data to run our analysis.
[3]https://www.nwea.org/map-growth/.

regression techniques to rigorously assess these variables, along with other possible determinants influencing student academic performance.

II. METHODOLOGY

A. DATA CLEANING & EXPLORATION[4]

During data exploration, we encountered missing data and irrelevant variables for our regression model. We eliminated survey columns that exhibited substantial missing values, along with other variables that were found to be unsuitable. This decision was informed by analyzing each variable's correlation with the outcomes using data visualization tools such as Tableau, as well as considering factors such as noise, ambiguity, and trends present in the datasets. Some remaining variables still had missing values; in these cases, we imputed data by making logical assumptions based on related information. For example, missing data on fathers' weekly work hours was imputed with zero for unemployed fathers or the dataset median for others. Categorical survey responses were numerically coded in Python, converting binary variables like gender into 0s and 1s and others.

B. DATA MODELLING

The model used to analyze factors influencing academic achievement (AA) in Beloit School District is represented by the following equation:

$$\mathbf{AA_i = \beta_0 + \beta_1 \, x \, (OSL)_1 + \beta_2 \, x \, (SE)_2 + \beta_3 \, x \, (EP)_3 + \beta_4 \, x \, (NY)_4 + \epsilon_i}$$

In this model, *AA* signifies academic achievement as the outcome variable, while *OSL* (Old School Label) represents whether a student attended an older or under-resourced school. SESESE denotes participation in special education programs, *EP* captures English proficiency levels, and *NY* reflects the number of years a family has lived in Beloit. The error term ε_i accounts for unexplained variance in academic achievement not captured by the included predictors. This model enables an assessment of the extent to which these variables individually and collectively impact academic outcomes in the district.

---

[4] *Google Colab*. (2019). Google.com. https://colab.research.google.com/drive/1DbCw3IUhvMoS8MR1_GO-mVdk_vR5bTy2?usp=sharing.

Professor Phan's dataset contained 106 variables, including students' demographic information (e.g., sex, race) and test scores, along with parental survey responses. The dataset tracked students' test scores from kindergarten onwards, but since the survey data was only collected in the latest testing year, we used only that year's scores as the regression target variable. To create a single measure for academic achievement, we averaged students' math and reading scores from both fall and winter as our target variable. Our feature selection process focused on identifying trends within each variable by generating visualizations in Python and Tableau. For instance, we examined Math/Read Test Scores by English Proficiency Levels to identify patterns, as shown in Figure 1, which revealed a trend in test performance by Eng Proficiency Levels. We systematically explored variables for meaningful trends in both reading and math scores.

Figure 1: Average Test Scores for English Proficiency Levels

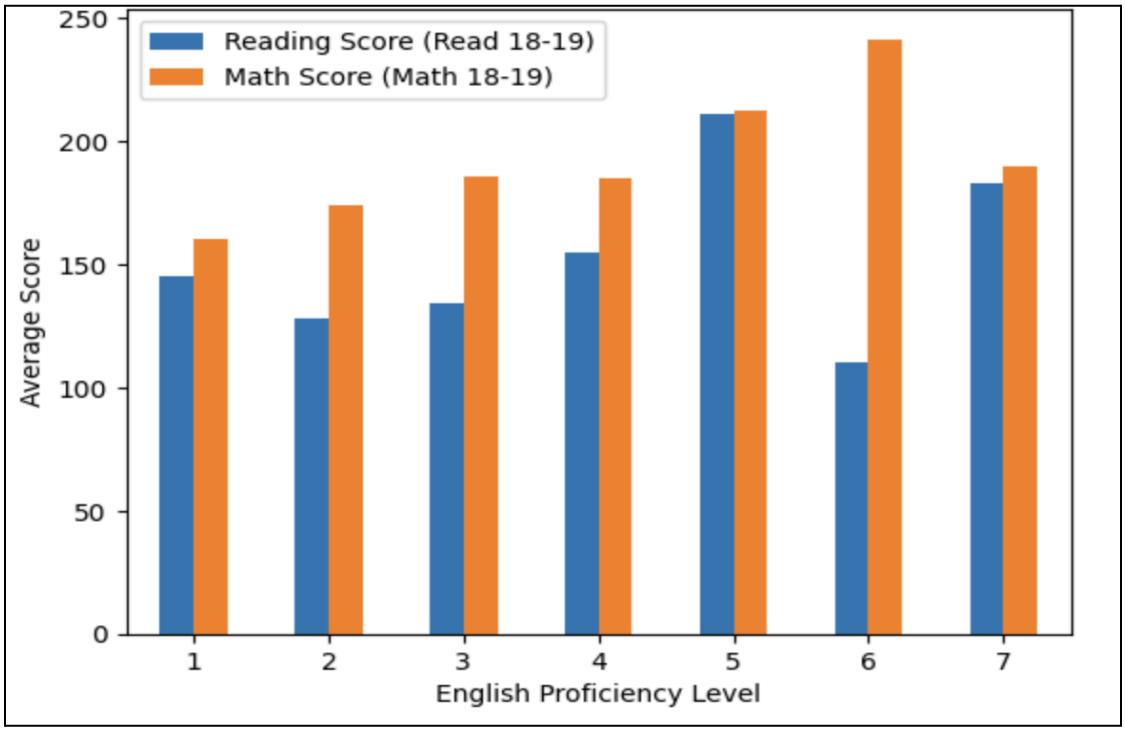
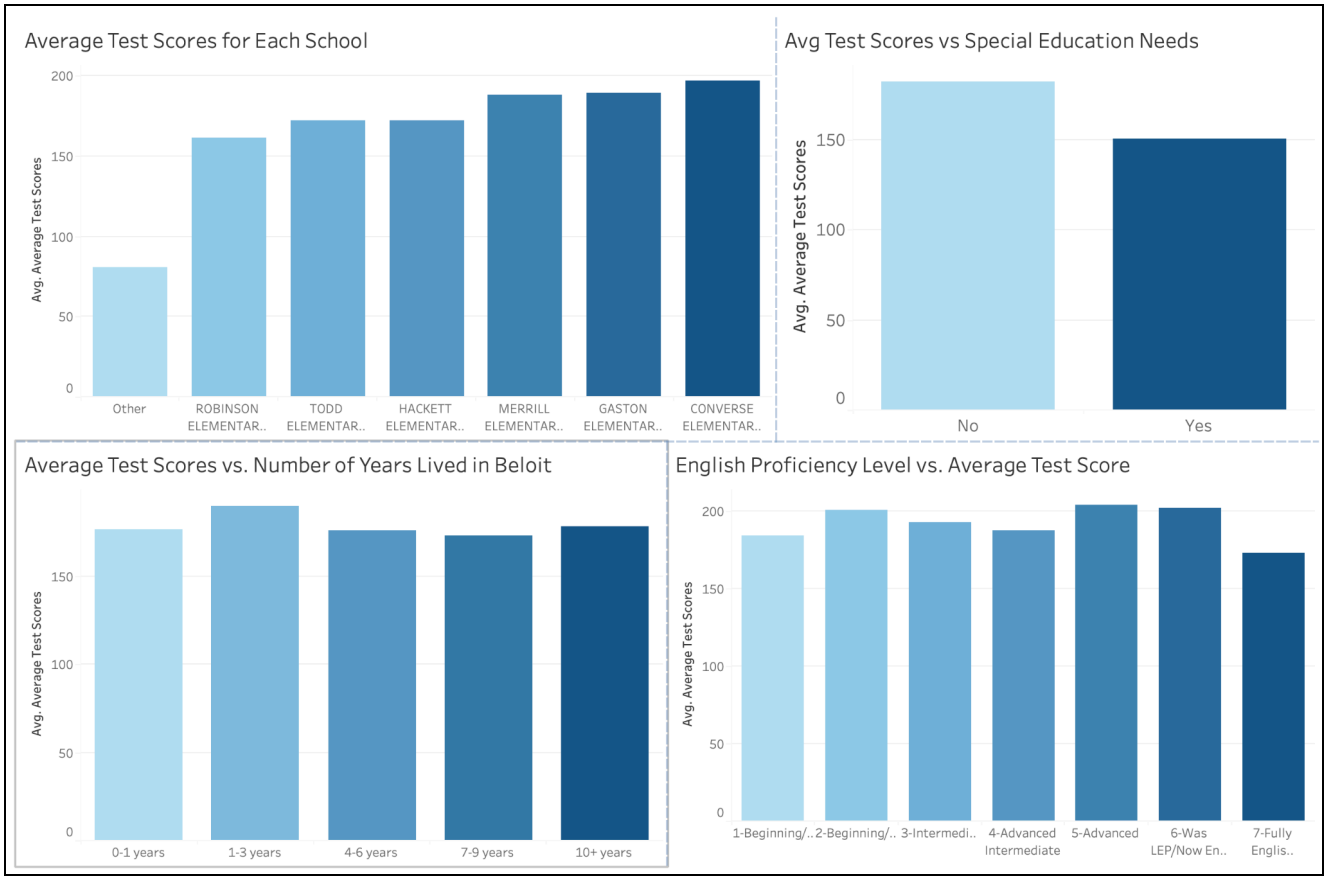


Figure 2: Data Visualization of our predictors



III. RESULTS & ANALYSIS

Figure 3: Line of Best Fit for Regression Analysis

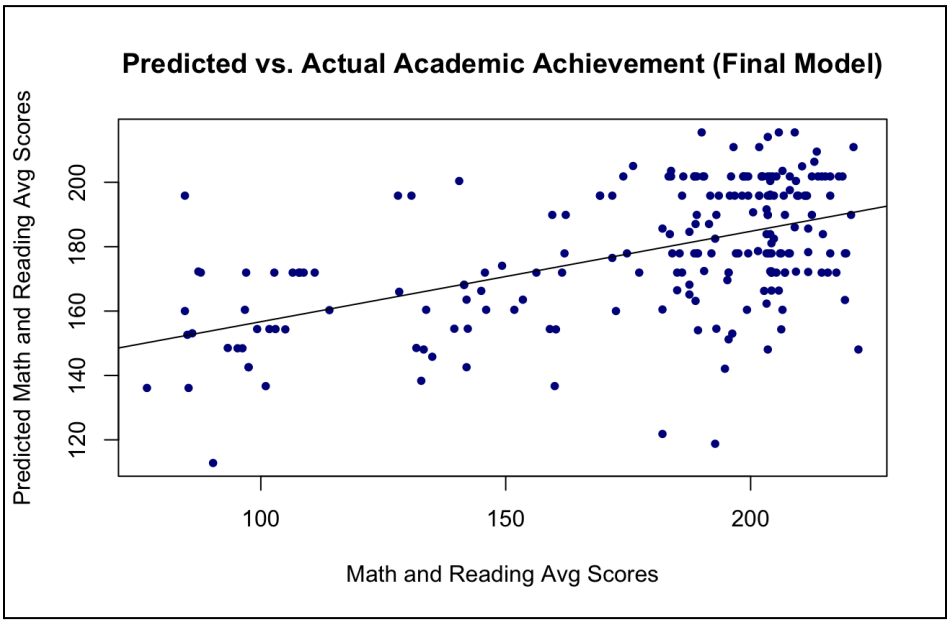


Table I: Linear Regression Data for the Academic Achievement Variables

Variables	Estimated Std.Dev	Std. Error	p-value	t-value
Intercepts	143.081	9.848	14.529	$< 2 \times 10^{-16}$ ***
Old School Labels (OLS)	5.974	1.238	4.825	2.70×10^{-6} ***
Special Education (SE)	-29.854	7.491	-3.985	9.32×10^{-5} ***
English Proficiency (EP)	5.876	1.162	5.055	9.41×10^{-7} ***
# of years stayed in Beloit (NY)	-4.561	2.091	-2.181	0.0303 *

Notes: This table reports the outcome for the Linear Regression as represented below, $AA_i = 143.081 + 5.974 \times (OSL)_1 - 29.854 \times (SE)_2 + 5.876 \times (EP)_3 - 4.561 \times (NY)_4 + Є_i$. Here, each coefficient reflects the influence of the predictor variables on academic performance, with positive or negative values indicating the direction of the relationship. *"OLS"* and *"EP"* are associated with positive coefficients, suggesting a positive effect on academic achievement, whereas "*SE*" and "*NY*" display negative coefficients, indicating a negative impact. The significance levels are denoted by asterisks, where ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$. These indicators highlight the likelihood that the observed relationships are not due to chance, with smaller p-values suggesting stronger evidence against the null hypothesis. The model's residual standard error is 33.13, based on 208 degrees of freedom, with an R-squared of 0.2808 and an adjusted R-squared of 0.267. The F-statistic of 20.3 ($p < 0.001$) indicates that the model significantly explains the variability in academic achievement across the variables considered.

IV. CONCLUSION AND RECOMMENDATION

Based on our analysis, policy interventions in the Beloit School District should prioritize enhanced English language support, equitable resource allocation for under-resourced schools, and community integration programs. Our regression model indicates that English proficiency is a significant positive predictor of academic achievement, suggesting that non-native English speakers may benefit from targeted support programs such as ESL (English as a Second Language) classes and additional bilingual resources. The observed positive coefficient for English proficiency underscores the importance of these initiatives in bridging achievement gaps. Additionally, students attending older, under-resourced schools displayed lower academic performance, highlighting a need for focused resource allocation to improve facilities, update instructional materials, and increase access to modern educational tools in these schools. Addressing these disparities would likely equalize educational opportunities and improve outcomes across the district. Furthermore, the model reveals a negative correlation between academic achievement and the number of years families have lived in Beloit, with newer

residents often experiencing challenges in adapting to the educational environment. Establishing community integration programs, such as newcomer orientation sessions and student mentorship initiatives, could facilitate a smoother transition for families and bolster student success. These policy recommendations, driven by the significant variables identified in our model, offer actionable strategies to enhance equity and educational attainment within the district.

V. FUTURE WORK

To strengthen the model and enhance predictive accuracy, future research should incorporate a wider set of control variables and utilize advanced modeling techniques. While our initial linear regression model identified English proficiency, school designation, special education status, and years lived in Beloit as key predictors of academic performance, adding variables such as household income, parental education, and neighborhood resources may provide a more comprehensive view of factors influencing outcomes.

Following data cleaning and initial analysis, we used linear regression to assess feature significance, which was then validated with advanced models like XGBoost and CatBoost. These techniques confirmed the importance of our selected variables, yet future studies could explore ensemble models or neural networks to capture potential nonlinear relationships. Expanding control variables and leveraging sophisticated algorithms will improve model accuracy and provide deeper insights into educational outcomes in the Beloit School District.