Improving Quality of Education Using Data Analytics in Chicago

Introduction

The objective of this research is to explore and understand the key factors influencing school performance within the Chicago Public Schools system, with a focus on instructional quality and student attendance. By analyzing metrics such as safety, environment, and academic achievement, this study aims to identify significant predictors of school success and address the disparities in performance across different regions of Chicago. The goal is to uncover actionable insights that can inform future educational policies and interventions, ensuring that all students, regardless of their geographic location or socioeconomic background, have access to high-quality education.

Dataset Description

Data Source: The dataset is from Chicago Data Portal (https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t/data_prev)

The dataset used in this study is a comprehensive collection of performance metrics from Chicago Public Schools for the academic years 2011-2012. The dataset consists of **566 rows** and **79 columns**. It includes detailed information on a wide range of variables, such as safety scores, environment scores, instructional scores, and various academic achievement indicators like literacy and math percentages across different grade levels. Additionally, the dataset includes demographic information and location data, allowing for a spatial analysis of school performance.

Methodology

The methodology employed in this research can be divided into several key stages:

1. Data preprocessing
2. Exploratory data analysis (EDA)
3. Feature selection
4. Linear regression models
5. Clustering

Data preprocessing involved cleaning the dataset to ensure it was suitable for analysis. This included handling missing values, converting categorical variables into numerical formats where necessary, and extracting location data for geographic analysis. The EDA phase involved a thorough examination of the dataset to identify patterns, outliers, and correlations between different variables. This stage was crucial for understanding the underlying structure of the data and informing the subsequent steps in the analysis.

Feature selection was performed to identify the most relevant variables for the linear regression models. Variables such as student attendance, environment score, and safety score were selected based on their presumed impact on instructional quality. This step ensured that the regression models were built on a solid foundation of meaningful predictors, improving the accuracy and interpretability of the results.

Linear Regression Models

1. The regression analysis reveals that both "Average Student Attendance" and "Environment Score" are significant predictors of "Instruction Score" in schools, explaining approximately **73.5%** of its variance. This **high R-squared** value suggests that the model effectively captures the relationship between these variables and instructional quality. Specifically, the positive coefficients for both predictors indicate that better attendance rates and a more supportive school environment are associated

with higher instruction scores. The strong impact of "Average Student Attendance," with a **coefficient slightly above 1**, underscores the critical role that consistent student participation plays in the effectiveness of instruction.

Moreover, the positive influence of "Environment Score" highlights the importance of the broader school environment in facilitating high-quality instruction. This finding suggests that efforts to improve the physical, social, and academic environment of schools can lead to meaningful improvements in educational outcomes. Overall, the results suggest that to enhance instruction quality, schools should prioritize strategies that improve student attendance and invest in creating a positive and supportive school environment.

2. The regression analysis presented investigates the impact of various family-related factors—namely, Family Involvement Score, Parent Engagement Score, and Parent Environment Score—on Average Student Attendance. The model yields an **R-squared value of 0.304**, which suggests that these variables collectively explain about **30.4% of the variance** in student attendance rates. While this indicates that family factors do influence attendance, a significant portion of the variation in attendance remains unexplained by the model, implying that other factors not included in this analysis may also play a substantial role.

   The coefficients for the Parent Engagement Score and Family Involvement Score are positive, indicating that higher scores in these areas are associated with higher student attendance, albeit the effects are relatively small. Conversely, the Parent Environment Score has a slightly negative coefficient, suggesting a small inverse relationship with attendance, which might be counterintuitive and could warrant further investigation to understand this dynamic better. Overall, the model highlights that while family engagement and involvement have a positive impact on student attendance, these factors alone are not strong predictors, suggesting the need to explore additional variables that could better explain student attendance rates.

Clustering

The clustering of schools on the map based on the criteria of being "**Healthy Schools Certified**" and having a "**Safety Score** greater than **70**" reveals some significant geographic patterns. Schools that are both certified as healthy and have high safety scores are concentrated in particular areas, suggesting that certain neighborhoods may have better access to resources, funding, or policies that promote healthier and safer school environments. For instance, it is noticeable that these schools are clustered in specific parts of the city, which could indicate a disparity in how different communities are served.

Moreover, the areas with a higher density of certified and safe schools may reflect neighborhoods that benefit from stronger community engagement or local government support. This clustering suggests that these regions might be more affluent or have more active involvement from parents and local stakeholders in ensuring the well-being of students. Conversely, the absence of such clusters in other regions could highlight areas where interventions are needed to improve school conditions, ensuring all students have equal access to safe and healthy learning environments. These insights could inform policymakers and educational administrators about where to focus their efforts and resources to promote equity across the school system.

Results

The results of the linear regression models and clustering analysis provide a comprehensive understanding of the factors that influence school performance in Chicago. The strong relationship between student attendance, environment score, and instructional quality underscores the importance of creating supportive school environments and promoting consistent attendance to enhance educational outcomes.

The clustering analysis further reveals geographic disparities in school performance, with certain areas of Chicago benefiting from higher concentrations of certified and safe schools. These findings suggest that while some regions have successfully implemented policies that promote healthy and safe learning environments, others lag behind, highlighting the need for more equitable resource allocation and support.

Conclusion

In analyzing the Chicago Public Schools dataset, a few key insights emerge that could be critical for educators, policymakers, and community stakeholders. One of the most significant findings is the strong relationship between student attendance, school environment, and instructional quality. Schools with higher student attendance rates and a supportive environment tend to have better instructional outcomes, as evidenced by higher instruction scores. This suggests that any efforts to improve educational outcomes in Chicago should prioritize policies that encourage consistent student attendance and create a positive, engaging school environment. The data underscores the importance of these factors in fostering effective teaching and learning, which are crucial to the overall success of students.

Additionally, the analysis reveals that while family involvement and engagement play a role in student attendance, they are not the sole determinants. The relatively low explanatory power of family-related factors on attendance rates indicates that other variables—such as socioeconomic status, community support, or school resources—might also be at play. This finding suggests that while it's important to engage families in the educational process, more comprehensive strategies are needed to address the broader issues that influence student attendance. Schools, therefore, might benefit from a more holistic approach that not only involves parents but also considers the broader community and systemic factors that affect student participation.

Finally, the geographic distribution of schools based on their certification as "Healthy Schools" and their safety scores reveals some disparities that are worth noting. Schools that are both certified and have high safety scores tend to be clustered in certain areas, suggesting that some neighborhoods are better served than others. This disparity raises important questions about equity in resource allocation and the need for targeted interventions in underserved areas. By identifying these clusters, the analysis provides a clear direction for where resources and efforts could be focused to ensure that all students, regardless of their location, have access to a safe and healthy learning environment. This geographic insight can be a valuable tool for guiding future policy decisions aimed at closing the gap between different communities within the city.

Future Work

- **Expand Analysis**: Future research should incorporate additional variables that may influence school performance, such as socioeconomic status, school funding, and community engagement, to provide a more comprehensive understanding.

- **Longitudinal Studies**: Conduct longitudinal studies to track how these factors evolve over time and assess their long-term impact on educational outcomes.

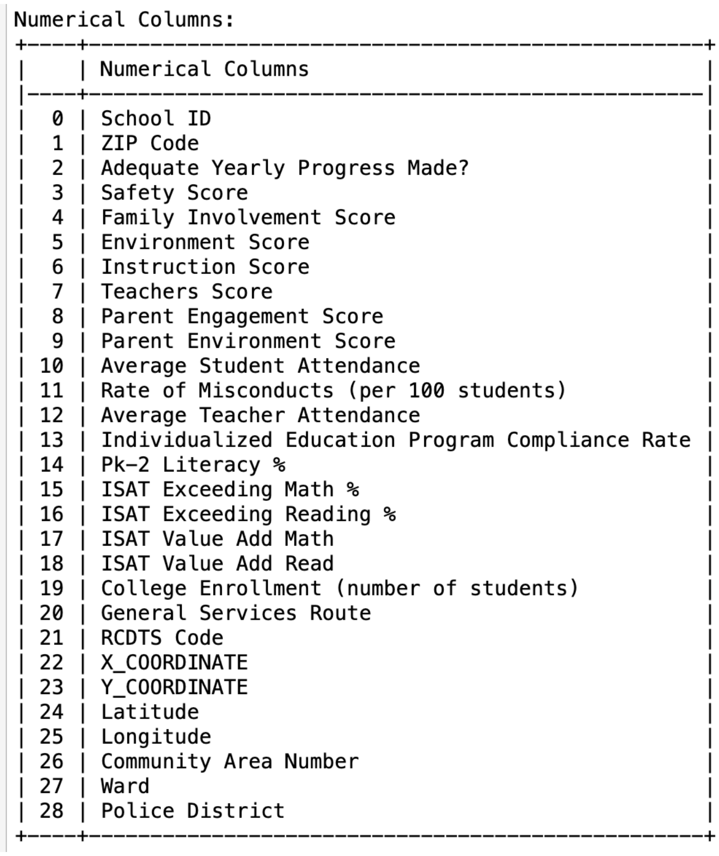
- **Geographic Disparities**: Further explore the geographic disparities identified in the clustering analysis to inform targeted interventions in underserved areas, enabling more effective strategies to promote educational equity.

- **Predictive Modeling**: Utilize predictive modeling techniques to explore the potential impact of policy changes or new initiatives on school performance, offering valuable guidance for decision-makers aiming to improve educational outcomes in Chicago and beyond.

References

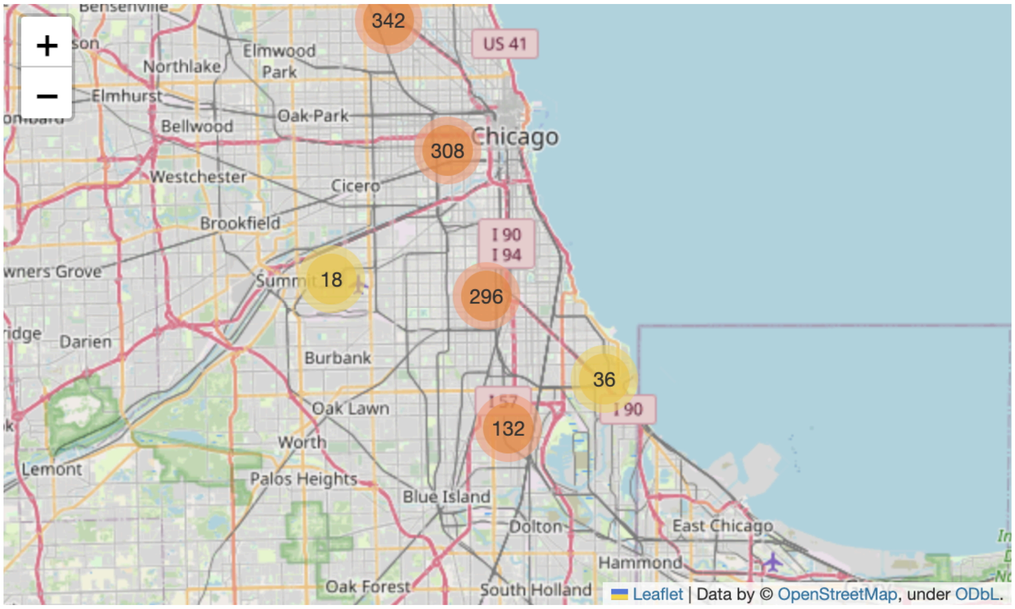
- 2021 Monthly school survey enrollment regression analysis ... Accessed August 12, 2024. https://ies.ed.gov/schoolsurvey/mss-report/supporting_files/mssregressionsummary.pdf
- Regression analysis: Graduation rate in Kentucky public ... Accessed August 12, 2024. https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1903&context=stu_hon_theses

Appendix



```
Numerical Columns:
+----+--------------------------------------------------+
|    | Numerical Columns                                |
|----+--------------------------------------------------|
|  0 | School ID                                        |
|  1 | ZIP Code                                         |
|  2 | Adequate Yearly Progress Made?                   |
|  3 | Safety Score                                     |
|  4 | Family Involvement Score                         |
|  5 | Environment Score                                |
|  6 | Instruction Score                                |
|  7 | Teachers Score                                   |
|  8 | Parent Engagement Score                          |
|  9 | Parent Environment Score                         |
| 10 | Average Student Attendance                       |
| 11 | Rate of Misconducts (per 100 students)           |
| 12 | Average Teacher Attendance                       |
| 13 | Individualized Education Program Compliance Rate |
| 14 | Pk-2 Literacy %                                  |
| 15 | ISAT Exceeding Math %                            |
| 16 | ISAT Exceeding Reading %                         |
| 17 | ISAT Value Add Math                              |
| 18 | ISAT Value Add Read                              |
| 19 | College Enrollment (number of students)          |
| 20 | General Services Route                           |
| 21 | RCDTS Code                                       |
| 22 | X_COORDINATE                                     |
| 23 | Y_COORDINATE                                     |
| 24 | Latitude                                         |
| 25 | Longitude                                        |
| 26 | Community Area Number                            |
| 27 | Ward                                             |
| 28 | Police District                                  |
+----+--------------------------------------------------+
```

Figure 1: List of Numerical Columns to be used later for linear regression



Map 1: Geographical distribution of Schools in Chicago

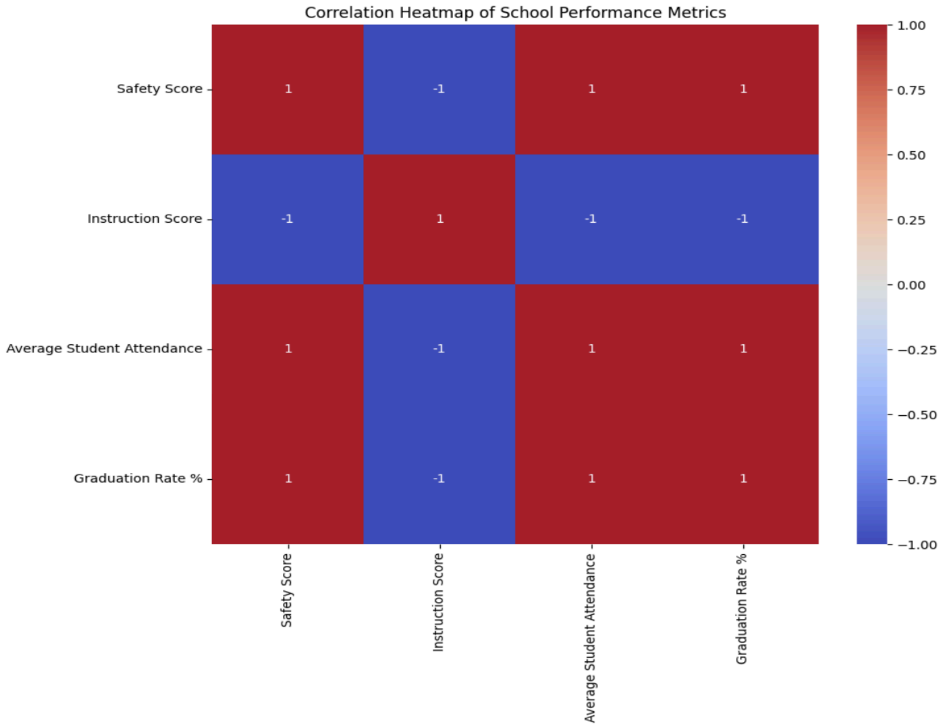


Figure 2: Correlation Heatmap of School Performance Metrics

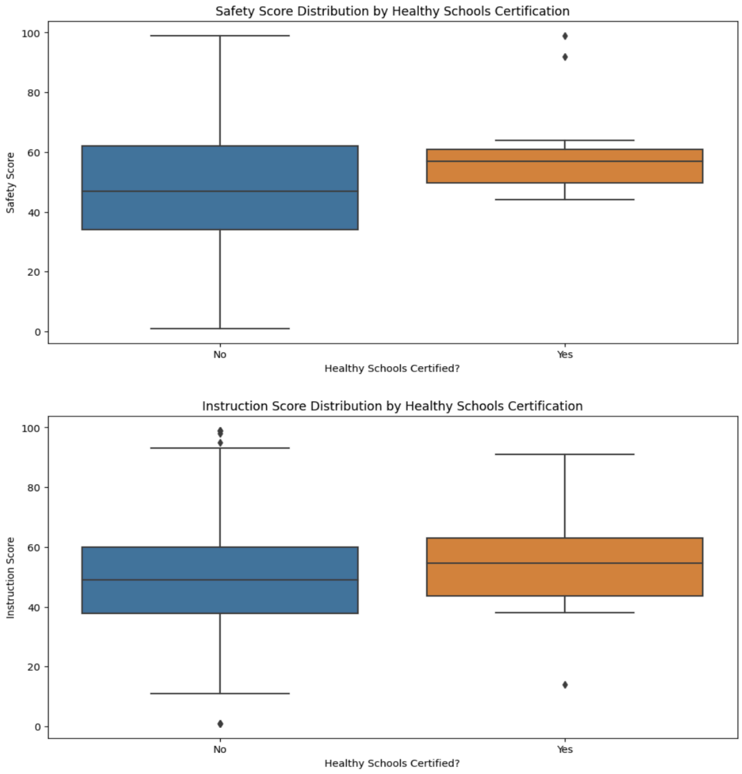
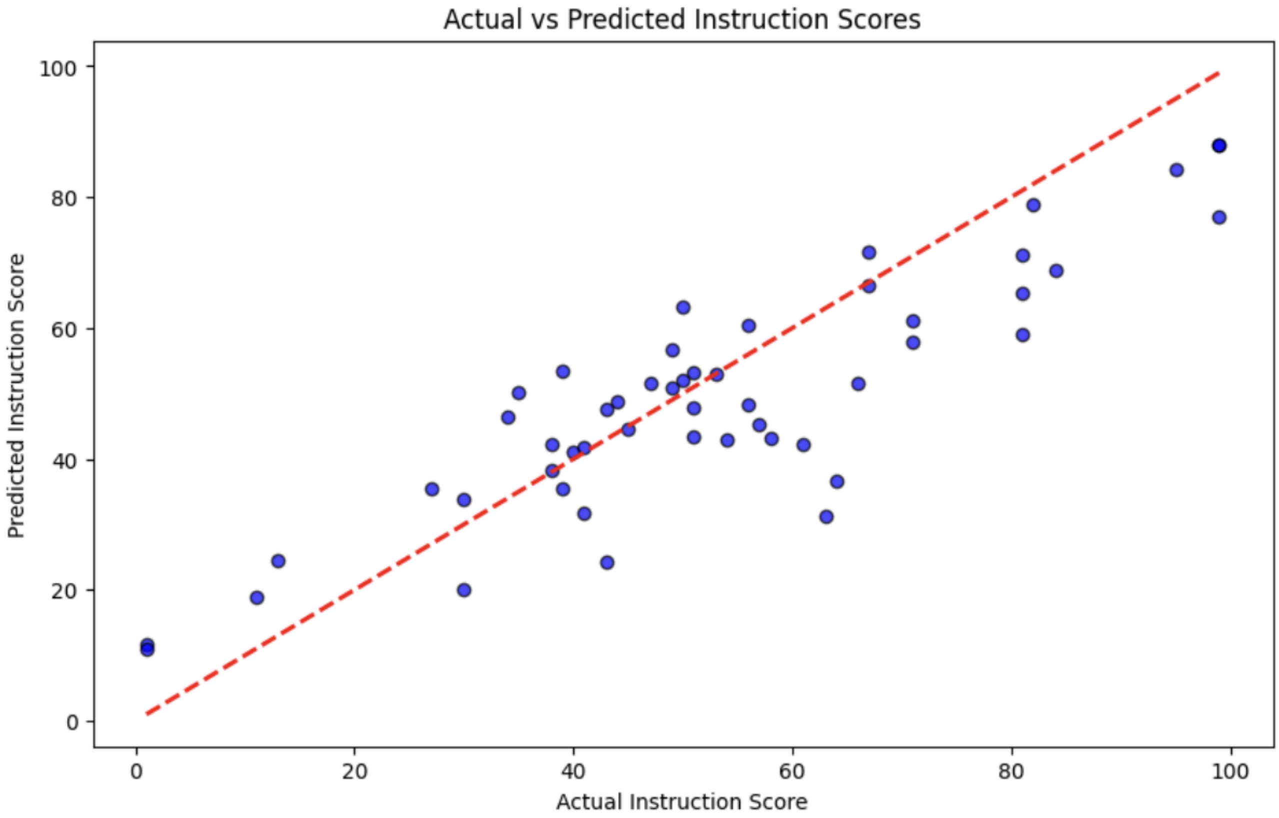
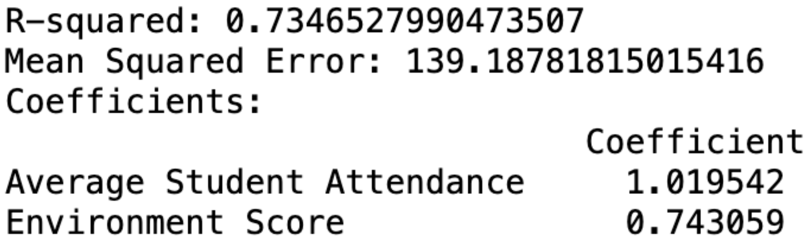


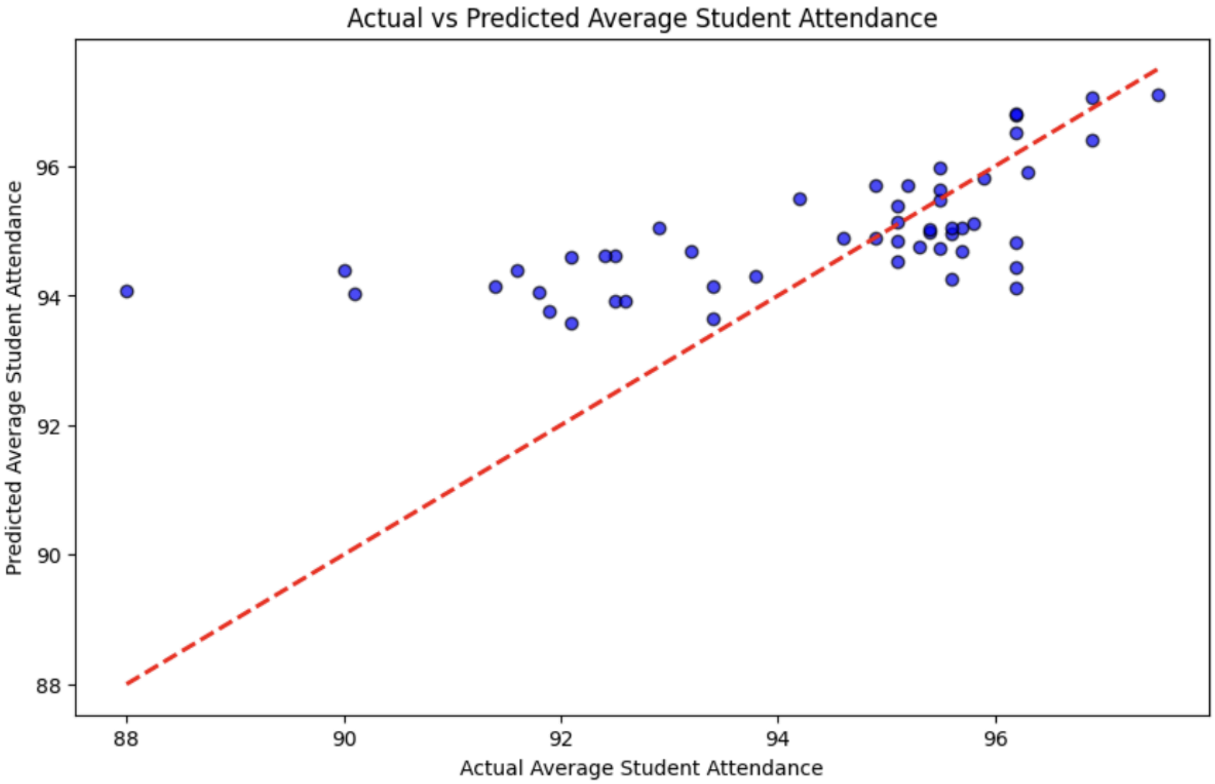
Figure 3: Boxplot of Safety Score and Instruction Score VS Healthy Schools comparison

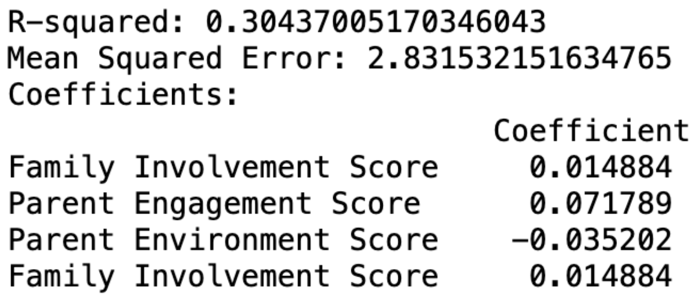




```
R-squared: 0.7346527990473507
Mean Squared Error: 139.18781815015416
Coefficients:
                             Coefficient
Average Student Attendance      1.019542
Environment Score               0.743059
```

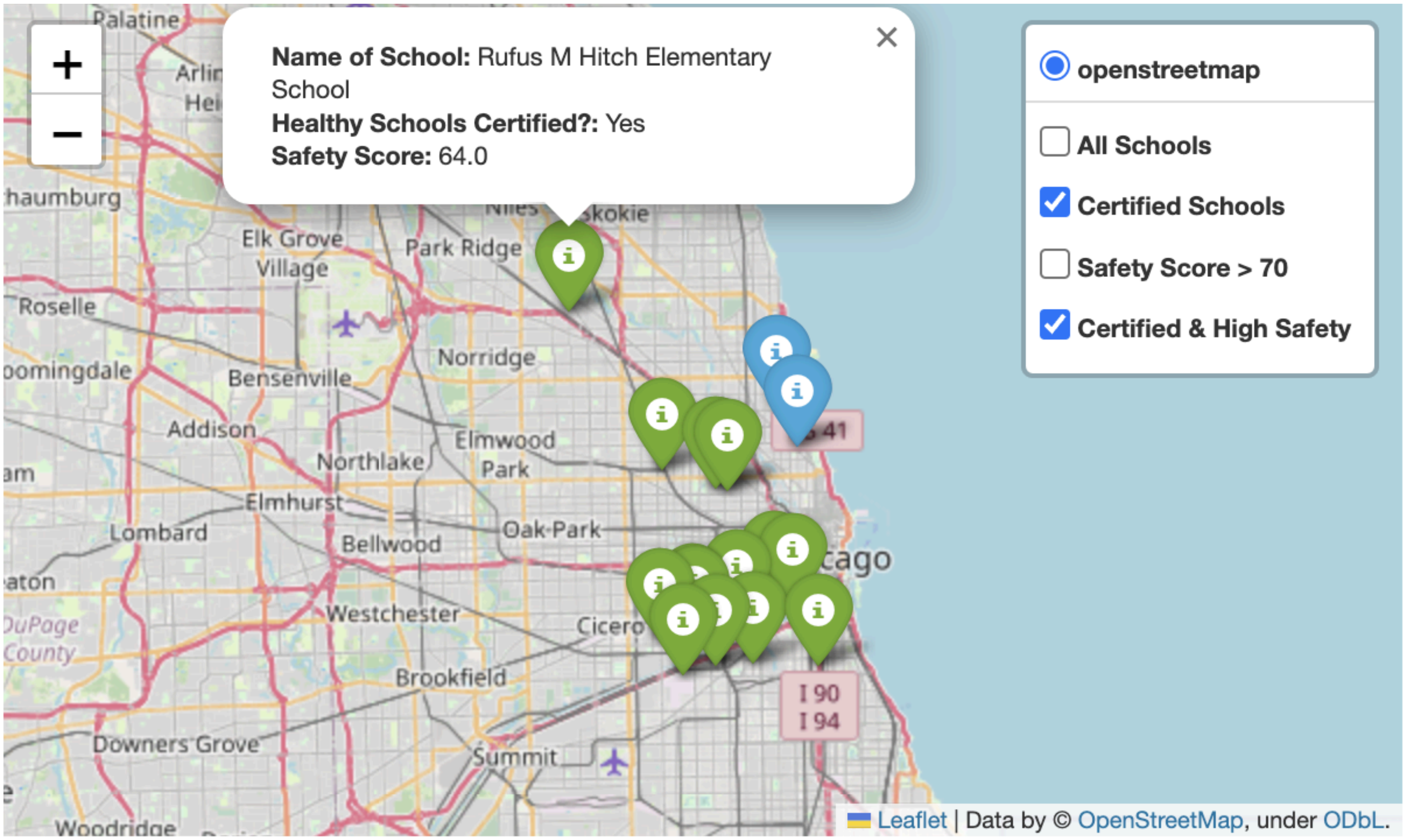
Figures 4 and 5: Actual vs Predicted Instruction Scores with evaluation of the model





```
R-squared: 0.30437005170346043
Mean Squared Error: 2.831532151634765
Coefficients:
                           Coefficient
Family Involvement Score     0.014884
Parent Engagement Score      0.071789
Parent Environment Score    -0.035202
Family Involvement Score     0.014884
```

Figures 5 and 6: Actual vs Predicted Average Student Performance with evaluation of the model



Name of School: Rufus M Hitch Elementary School
Healthy Schools Certified?: Yes
Safety Score: 64.0

● openstreetmap

☐ All Schools
☑ Certified Schools
☐ Safety Score > 70
☑ Certified & High Safety

Map 2: Folium Python map generated with clustering on Health Schools and Higher Safety as filters