The Relationship Between Iowa Public School Budgets and Student Proficiency

Nicolas J. Vetter

Northwest Missouri State University, Maryville MO 64468, USA
S567397@nwmissouri.edu or nvetter73@gmail.com

Abstract. This report examines the relationship between Iowa public school budgets/district size, and student proficiency by using SQL, exploratory data analysis (EDA), and machine learning in Python. The EDA revealed that most schools spend around 8,000 dollars per pupil and are around 80 percent reading proficiency. Also, smaller schools tended to have a higher average proficiency, however, this was because of the higher amount of small-sized schools. Despite these observations, the correlation between school expenditure/district size and student proficiency was poor.

In section 5, models including Linear Regression, Logistic Regression, Random Forests, and Decision Trees were run to explore the effect of budget/enrollment on proficiency. However, the results were inconclusive due to the small dataset size (300 instances) and the lack of patterns. None of the models were able to show predictive power or correlation between expenditure/enrollment, and reading proficiency. The report's findings show that factors beyond school budgets and size influence student success. Because of this, more variables should be considered to find a correlation. Overall, the report shows the limitations of budget and enrollment data in predicting student performance. It also emphasizes the need for further analysis of other factors that show educational success.

Keywords: Data Analytics · Iowa Public School · Budgets · Proficiency

Helpful Links:

- GitHub Capstone Repository:
  https://github.com/VetterNic2/msda-capstone

- Public Overleaf:
  https://www.overleaf.com/read/bhvmwnchhksz#8526a8

1 Introduction

Whether you grew up in the house of a teacher/parent or not, you probably have a memory or two about your high school life. Whether your school was rich or poor, all Iowa Public Schools have a duty to give their students the best chance to succeed after they graduate. In order to give their students the opportunity to be successful, the school system needs support their pupils. In state legislation, there is an amount a public school should be funded per pupil. As referenced in a California expenditure report here: (1). This project is going to show which Iowa Public Schools invested the most/least in their students and whether district size affects reading proficiency. This report could give the reader an idea of how much of an investment a school should spend to make their student the most successful they can be, and whether there is a correlation between reading proficiency and investment per pupil/district size. Success will be measured by proficiency scores in this report. This data will be cleaned in PostgreSQL, analyzed/visualized with EDA and machine learning in Python, along with a final visual in Tableau. The data sources and references are shown in their respective sections.

Limitation: Iowa Public Schools have several different budgets for all kinds of expenditures. Because of the short report time frame, every budget within the Iowa Public School system will not be analyzed. This report will be kept simple and analyze the budget every Iowa Public school used in the 2017 academic year:

general instruction. However, there has been some solid research on the school expenditure vs. proficiency dynamic in the past. The following reference dives into a few more external factors that will not be reported on in this project:(4).

1.1 Goal of this Project

This report will determine if expenditure per pupil and district enrollment size affect the average reading proficiency of a school district. This goal will be attained by cleaning and validating the data, conducting exploratory data analysis, and running various machine learning models to determine if reading proficiency can be predicted with the independent variables of expense and district size.

2 Data Sources

- Math and Reading Proficiency in Iowa by School Year: https://data.iowa.gov/Primary-Secondary-Ed/Math-And-Reading-Proficiency-in-Iowa-by-School-Yea/f3h8-mnxi/about_data

- Iowa School District Expenditures by Fiscal Year: https://data.iowa.gov/School-Finance/Iowa-School-District-Expenditures-by-Fiscal-Year/uutu-bzs3/about_data

2.1 Data Collection

- **Math and Reading Proficiency in Iowa by School Year:**

  This data set was collected from the Iowa.gov website as public information. The Iowa.gov website has an "Action Query" function that helps the writer filter the necessary data from the dataset. This is called preliminary cleaning of the dataset and the dataset will be cleaned again in section 3 of this report. The writer will only be using the reading proficiency rating of the 2017 11th-grade students in each respective district. This is to shorten the report for the tight time window.

- **Iowa School District Expenditures by Fiscal Year:**

  This data set was collected from the Iowa.gov website as public information. The Iowa.gov website has an "Action Query" function that helps the writer filter the necessary data from the dataset. This is called preliminary cleaning of the dataset and the dataset will be cleaned again in section 3 of this report. The writer will only be using one year's worth of data from the 2017 fiscal/academic year. This is to shorten the report for the tight time window.

2.2 Data Description

- **Math and Reading Proficiency in Iowa by School Year:**

  The total storage space of this structured dataset is 48KB. It contains 610 records and 14 attributes. However, the writer won't be using all of the 14 attributes in this report. Some were deemed irrelevant to the analysis and report. With that in mind, the attributes to be used for analysis are as follows with the datatype shown in parenthesis behind the attribute: School Year(Number), Topic(Text), Grade(Number), District(Text), District Name(Text), Percent Proficient(Number), Proficient Category(Text).

- **Iowa School District Expenditures by Fiscal Year:**

  The total storage space of this structured dataset is 366KB. It contains 3509 records and 13 attributes. However, the writer won't be using all of the 13 attributes in this report. Some were deemed irrelevant to

the analysis and report. With that in mind, the attributes to be used for analysis are as follows with the datatype shown in parenthesis behind the attribute: Year(Number), Dist(Text), District Name(Text), Fund(Text), Expenditures Per Pupil(Number), Amount(Number), Enrollment Category(Text), Enrollment Category Number(Number).

3 Data Cleaning/Manipulation using PostgreSQL

3.1 Cleaning Process

– **Process and Tools of Data Cleaning:**

The preprocessing of data was very minimal for these data sources. This is because of the lack of missing values within the CSV files. The only missing values were related to the geographic location of school districts, and for this report, that was deemed irrelevant. Because of the irrelevance, those missing values were dropped from the SQL tables. There were no relevant missing values because Iowa.gov/data pre-cleans data so it is accurate and ready for third-party analysis.

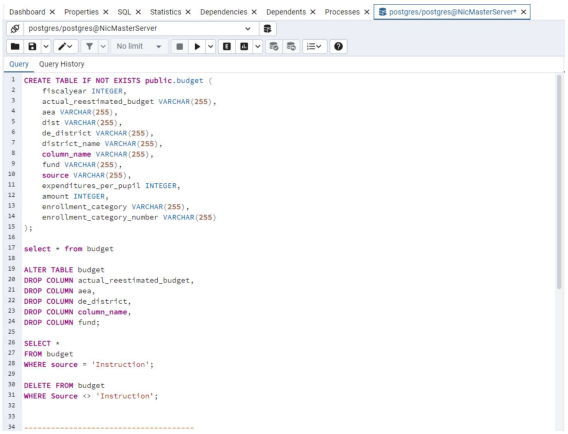
Since the data came from two different CSV files, SQL needed to be used to clean and join the two files together. Specifically, PgAdmin and PostgreSQL were selected because of the ease of table creation and import/exporting of CSV data. First, the tables needed to be created using the PgAdmin interface and copy statements for both CSV files. Second, the unnecessary attributes needed to be dropped from the tables to ensure simplicity and accuracy. Along with the unnecessary attributes, this report is only going to analyze reading proficiency. This decision was made because of the tight timeline of the report and the need for only one dependent variable. Third, the two revised tables needed to be joined with a JOIN statement. After the JOIN statement ran successfully, the new data table was saved to a CSV file for further analysis.

3.2 JOIN Statement and Figures

– **PostgreSQL JOIN:**

The two CSV files described in the previous section were joined to only show relevant attributes and information. This will make it much simpler to analyze if there is only one CSV file to run through the machine learning model and create visualizations. The JOIN statement can be seen at the bottom of Fig. 2.

– **Cleaning/Manipulation Figures**



```
CREATE TABLE IF NOT EXISTS public.budget (
    fiscalyear INTEGER,
    actual_reestimated_budget VARCHAR(255),
    aea VARCHAR(255),
    dist VARCHAR(255),
    de_district VARCHAR(255),
    district_name VARCHAR(255),
    column_name VARCHAR(255),
    fund VARCHAR(255),
    source VARCHAR(255),
    expenditures_per_pupil INTEGER,
    amount INTEGER,
    enrollment_category VARCHAR(255),
    enrollment_category_number VARCHAR(255)
);

select * from budget

ALTER TABLE budget
DROP COLUMN actual_reestimated_budget,
DROP COLUMN aea,
DROP COLUMN de_district,
DROP COLUMN column_name,
DROP COLUMN fund;

SELECT *
FROM budget
WHERE source = 'Instruction';

DELETE FROM budget
WHERE Source <> 'Instruction';
```

Fig. 1. Create/Clean District Expenditures

```
CREATE TABLE school_performance (
    school_year INTEGER,
    topic VARCHAR(50),
    grade INTEGER,
    district_id VARCHAR(10),
    district_name VARCHAR(100),
    proficient INTEGER,
    total INTEGER,
    percent_proficient NUMERIC(5, 2),
    proficient_category VARCHAR(50),
    district_office_location VARCHAR(100),
    iowa_zip_code_tabulation_areas VARCHAR(100),
    iowa_watersheds_huc10 VARCHAR(100),
    iowa_watershed_subbasins_huc08 VARCHAR(100),
    us_counties VARCHAR(100)
);
ALTER TABLE school_performance RENAME TO performance;

ALTER TABLE performance
DROP COLUMN district_office_location,
DROP COLUMN iowa_zip_code_tabulation_areas,
DROP COLUMN iowa_watersheds_huc10,
DROP COLUMN iowa_watershed_subbasins_huc08,
DROP COLUMN us_counties,
DROP COLUMN school_year,
DROP COLUMN grade;

DELETE FROM performance
WHERE topic= 'Math';

JOIN:

SELECT b.*, p.topic, p.proficient, p.total, p.percent_proficient, p.proficient_category
FROM budget b
JOIN performance p ON b.dist = p.district_id
WHERE b.fiscalYear = 2017;
```

Fig. 2. Create/Clean Proficiency and JOIN

3.3 Clean Data Overview

– **Attributes and Variables:**

The attributes after the cleaning process was completed are as follows:
"fiscalyear","dist","district name","source","expenditures per pupil","amount","enrollment category","enrollment category number","topic","proficient","total","percent proficient","proficient category".

Below is a sample of the CSV's first line of the cleaned data:
2017,"0009","AGWSR","Instruction",7989,4997256,"600-999","3","Reading",32,41,"78.00","70.1 - 80".

If the data shown in the line above is surrounded by "", it is a string/text data type. However, if the data shown in the line above is surrounded by nothing, it is an integer data type.

To align with the goal of the report, the dependent variable of the project is "percent proficient" and the independent variables are all of the other attributes contained within the cleaned CSV file(outlined above). However, the main independent variable that will be analyzed is "expenditures per pupil".

4 Exploratory Data Analysis

4.1 EDA Process

– **Process and Tools of Exploratory Data Analysis:**

Choosing the right tools to conduct EDA is very important for any analyst. For the analysis of the Iowa Public School proficiency and expenditures, Jupyter Notebooks was the best option. This is because of the amazing visualization capabilities of Jupyter within a virtual environment. The modules used were as follows: pandas, matplotlib.pyplot, and seaborn. The Jupyter Notebook can also be seen at the GitHub Repository linked on the first page of the report.

This report's EDA process is outlined with the following sections: Data Acquisition - Read and View Dataset, Data Inspection - Data Type / Numerical Stats / Missing Value Verification, Numerical Attribute Histograms, Categorical Attribute Bar Charts, and Initial Visualizations. All of these sections provide insights into the data selected for this report/analysis. Also, there are some very interesting correlations, or lack of correlations, between certain attributes.

4.2 Data Acquisition - Read and View Dataset

– **Data Acquisition Overview:**

The first step of EDA is to acquire the cleaned data. The data was read from the CSV file using the pandas module. Also, the head of the dataset was shown as a confirmation that the data was read into the notebook. The code and head of CSV associated with this process can be found in Fig. 3 below.

– **Data Acquisition Figure**

Data Acquisition - Reading and Viewing Dataset

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import numpy as np


# Loading my CSV file
df = pd.read_csv('cleandata.csv')
print(df.head())
```

[4]

```
   fiscalyear  dist        district_name       source  expenditures_per_pupil  \
0        2017     9                AGWSR  Instruction                    7989
1        2017   441                AHSTW  Instruction                    7318
2        2017    27  Adel DeSoto Minburn  Instruction                    7174
3        2017    63      Akron Westfield  Instruction                    8339
4        2017    81               Albia  Instruction                    7048

     amount enrollment_category  enrollment_category_number    topic  \
0   4997256             600-999                           3  Reading
1   5744497             600-999                           3  Reading
2  11256504           1,000-2,499                         4  Reading
3   4302923             300-599                           2  Reading
4   8506189           1,000-2,499                         4  Reading

   proficient  total  percent_proficient proficient_category
0          32     41                78.0          70.1 - 80%
1          45     54                83.3          80.1 - 90%
2         105    116                90.5         90.1 - 100%
3          29     34                85.3          80.1 - 90%
4          63     77                81.8          80.1 - 90%
```

Fig. 3. Data Acquisition Process

4.3 Data Inspection - Type / Stats / Missing Values

– **Data Inspection Overview:**

The second step of EDA is to inspect the pandas dataset. The data was inspected for data type information, statistics for numerical columns, and finally, a missing value verification was conducted. However, no missing values were found, so no figures are shown in that notebook summary. The figures associated with this process can be found in Fig. 4-5 below.

– **Data Inspection Figures**

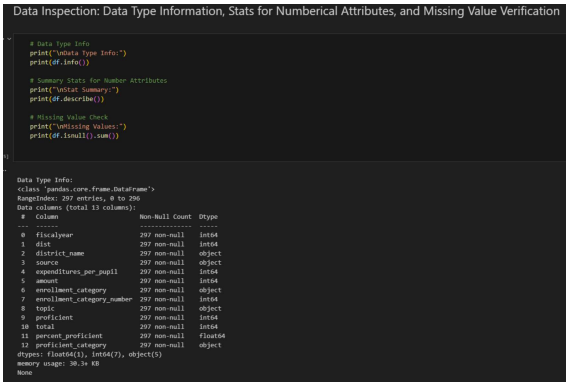


Fig. 4. Data Type information

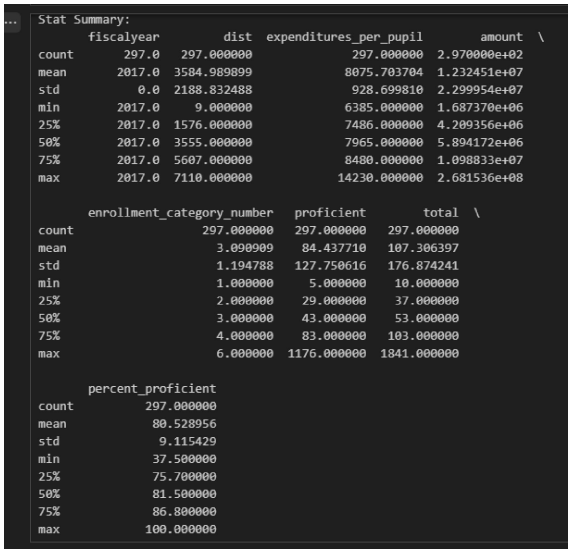


Fig. 5. Pandas Stats

4.4 Numerical Histograms

– **Numerical Histogram Overview:**

The third step of EDA is to conduct a data distribution of numerical attributes within the dataset. More specifically, the most interesting attributes were: "expenditure per pupil" and "percent proficient". While reviewing these histograms, the data concludes that most schools spend around 8,000 dollars per student and 80 percent reading proficiency in the 11th-grade class. The code and histograms associated with this process can be found in Fig. 6-8 below.

– **Numerical Histogram Figures**

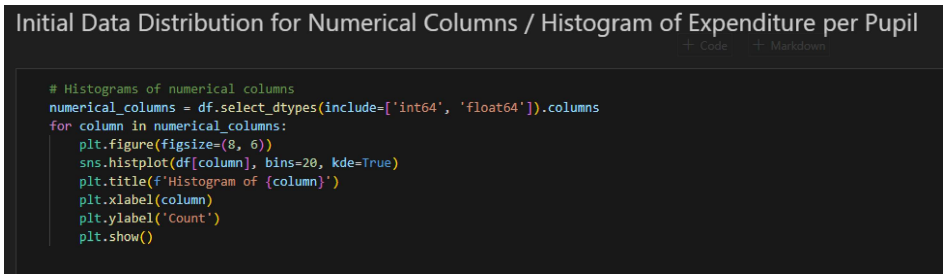


Fig. 6. Histogram Code

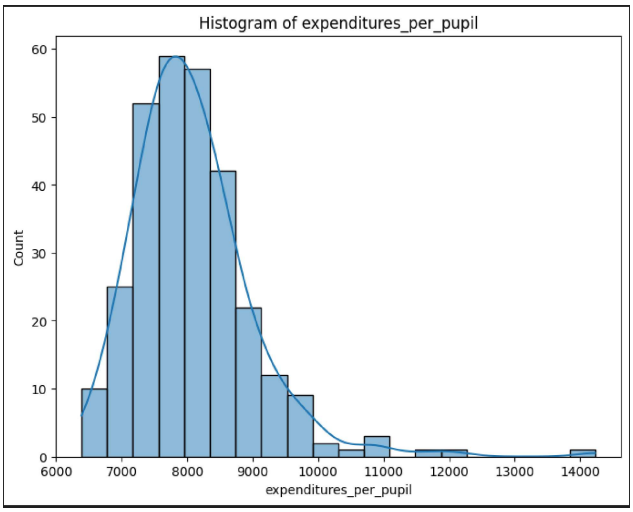


Fig. 7. Expenditure per Pupil Histogram showing Bell Curve around 8,000

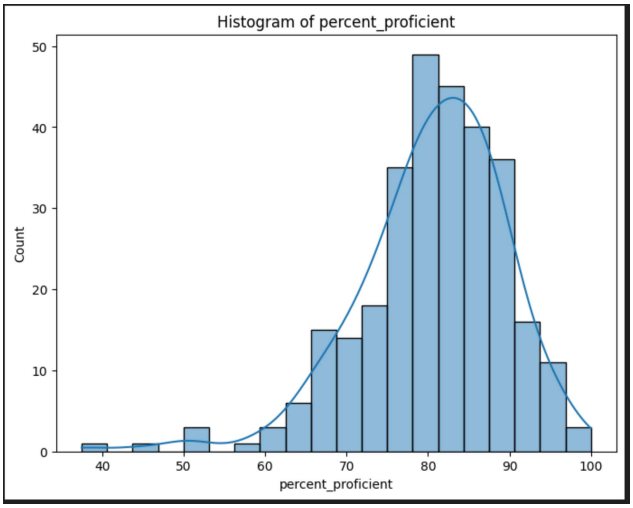


Fig. 8. Reading Proficiency Percentage Histogram showing Bell Curve around 80 percent

4.5 Categorical Bar Charts

– **Categorical Bar Chart Overview:**

The fourth step of EDA is to conduct a data distribution of categorical attributes within the dataset. More specifically, the most interesting text attributes were: "proficient category" and "enrollment category". While reviewing these bar charts, the data concludes that most schools are in the 70-100 percent proficient category in the 11th-grade class. Also, the enrollment category bar chart shows a large number of small schools compared to the larger schools in terms of enrollment. These two attributes are important because most small schools have a higher proficiency percentage. After all, the total number of students is so much smaller. The code and bar charts associated with this process can be found in Fig. 9-10 below.

– **Categorical Bar Chart Figures**

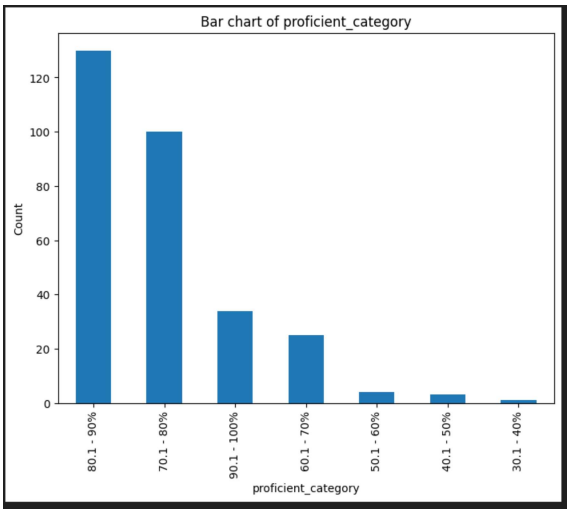


Fig. 9. Proficient Category Bar - most in 70-100 percent

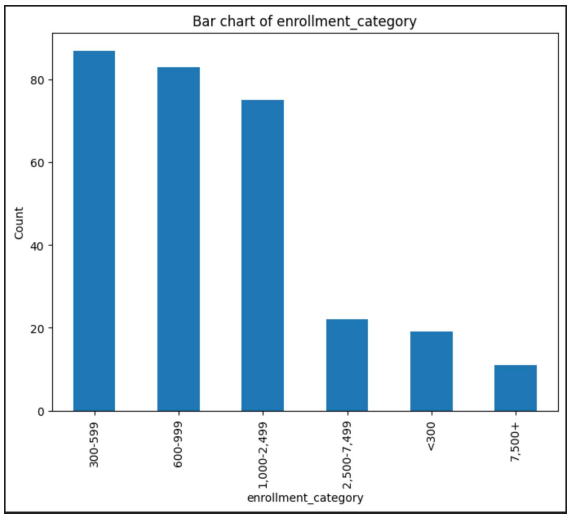


Fig. 10. Enrollment Category Bar - larger amount of smaller districts

4.6 EDA Visualizations

– **Visualization Overview:**

The fifth and final step of EDA is to create some visualizations that accurately capture the data, and check if any correlations need to be investigated further.

The first visualization was created to view the average percent proficient for each enrollment category number. The enrollment category number is determined by total number of enrolled students in the school district. Smaller enrollment category numbers are for small schools, and large schools have a higher enrollment category number. This line chart shows the highest average reading proficiency in each of the district size categories.

The second visualization was created so the minimum and maximum reading proficiency could be shown to the viewer. This is a very interesting visualization, and it captures the disparity of Iowa Public School systems.

The final visualization was created to show any preliminary correlation between budget expenditure and reading proficiency. There is a lot of good information to be gained from this scatter plot. First, it looks like most schools hang around the 8,000 dollars per pupil range. Also, the lower left quartile is pretty empty compared to the other quartiles. Finally, all of the schools spending more than 10,500 dollars per pupil are over 80 percent proficient in 11th-grade reading. The bottom/top 10 code block and visualizations associated with this section can be found in Fig. 11-14 below.

– **EDA Visualization Figures**

```
# Sort DataFrame by percent_proficient
df_sorted = df.sort_values(by='percent_proficient', ascending=False)

# Select top 10 and bottom 10 public school districts
top_bottom_10 = pd.concat([df_sorted.head(10), df_sorted.tail(10)])

# Creating a bar chart
plt.figure(figsize=(12, 8))
bars = plt.bar(top_bottom_10['district_name'], top_bottom_10['percent_proficient'], color='skyblue', edgecolor='black')

# Labeling title and axis
plt.title('Best and Worst of 2017 Reading Proficiency - Top and Bottom 10')
plt.xlabel('Iowa School District Name')
plt.ylabel('Percent Proficient (%)')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.grid(axis='y')  # Add grid only on y-axis for bar chart

# Labels
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval, round(yval, 1), va='bottom', ha='center')

# Show plot
plt.tight_layout()
plt.show()
```

Fig. 11. Code Sample of Top/Bottom 10

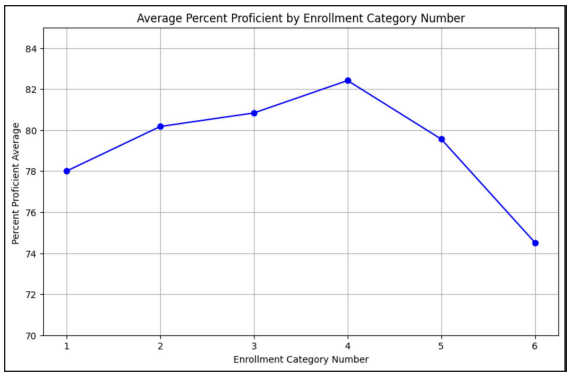


Fig. 12. Average Proficiency per District Size - district size 3-4 have highest reading proficiency

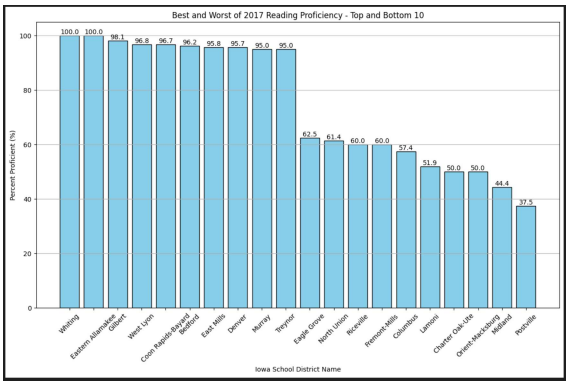


Fig. 13. Top/Bottom 10 in Reading Proficiency - the highs and lows of Iowa public education

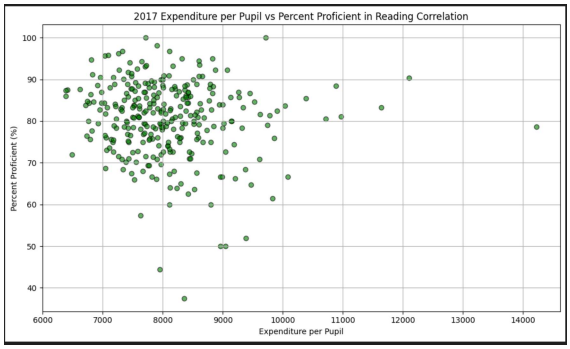


Fig. 14. Expense vs. Proficiency Correlation

4.7 Conclusion of EDA

– **Conclusions Drawn from EDA:**

The above statistics and visuals do a great job of exploring the data and set up how the machine-learning portion of the report will be attacked. The visuals do a great job of providing a landscape of the Iowa Public School systems as well.

Some of the categorical and numerical columns didn't need to be visualized, but it does give the viewer a great summary of the data set being used. Exploratory Data Analysis should give anyone great insight into the world of Iowa Public Schools and this step is one of the most crucial in all of Data Analytics.

With the results of this EDA in mind, we will focus on the correlation between enrollment/expenditure attributes, and how they affect the dependent variable of "percent proficient". Based on the EDA, it looks like the size of enrollment has a bigger effect on reading proficiency than expenditure per pupil, but this hypothesis will be experimented with in the machine learning section of the report.

5 Machine Learning in Python - Correlation

5.1 Machine Learning Overview

– **Process, Methods, and Implementation:**

Machine Learning has become standard practice in the world of Data Analysis. Even just a few years ago, the thought of teaching machines to predict outcomes sounds like something out of a Sci-fi film. Throughout this report, we have discussed the possible correlation of expenditure/size of a district with the target of Reading Proficiency. However, I have found that no models were fit for the data and there was no correlation between those three important features. The two machine learning models used were Linear and Logistic Regression with an 80-20 train-test split. Random Forest Models and Decision Trees were also used, but they were an even worse fit for the data somehow.

No models fit the data, as reflected by their error rates. Also, the predictability is 0, as reflected by the R-squared metric. Scaling was also attempted to reduce the Error rates and predictability, but no progress was shown in the fitting of the models. Also, feature engineering wasn't an option because of the lack of size of the dataset from previously thorough cleaning. The small size of the dataset only having 300 instances, also led to an inability to fit predictability with any machine learning models.

5.2 Model 1: Linear Regression

– **Linear Regression Overview:**

Linear Regression was used in a uni-variant nature, meaning the model only measured the correlation/predictability between the feature(expenditure per pupil or enrollment category) and target(reading proficiency). This was chosen to show the correlation of each feature independently. The results of the predictive analysis are as shown in Fig 15-16 below.

However, there are no complete failures in data analysis projects. The results of the linear regression simply show that there was no fit for the data in the linear regression model, and there is no predictability/correlation between the three features.

Both of the Linear Regression Models showed a relatively high mean squared error and a very low R squared value. This can be interpreted as the model's prediction errors are relatively high and the 1 percent R squared value only correlates with 1 percent of reading proficiency rates. Once again, scaling was attempted, but the results were not any better, and feature engineering was not feasible because of the extensive cleaning of the dataset and the small size.

– **Linear Regression Model Figures**

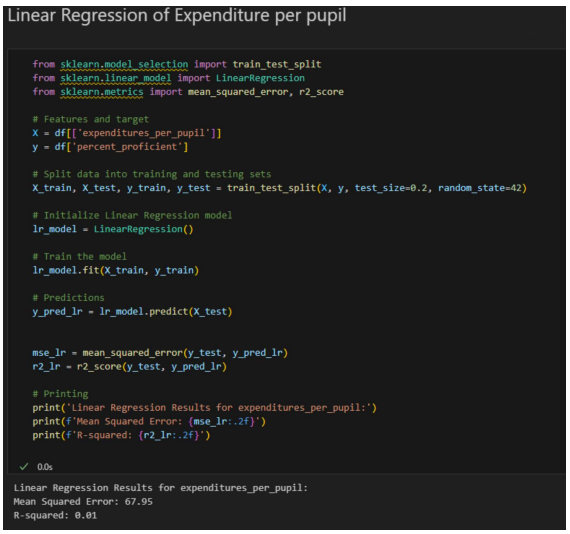


Fig. 15. A screenshot shows the machine learning model used to detect the correlation predictability between Expense per pupil and Reading Proficiency

Fig. 16. A screenshot shows the machine learning model used to detect the correlation predictability between enrollment size and Reading Proficiency

5.3 Model 2: Logistic Regression

– **Logistic Regression Overview:**

Logistic Regression is an odds-based prediction model. The main difference between Linear and Logistic is that Logistic is binary, while Linear is continuous [5]. The binary nature of Logistic Regression was chosen because Proficiency thresholds are extremely important for any district in the Iowa Public

School system. The reading proficiency threshold was determined to be 80 percent. This was selected because of the information gained from the exploratory data analysis.

The results of the Logistic Regression Model can be seen in Figure 17. For an in-depth review of the code, please see the GitHub link at the top of the report. Please know scaling was attempted for better model performance, but it was not helpful for this dataset and lack of correlation/predictability.

Throughout the interpretation of the results, class 0 and class 1 will be used to reference as follows: percent proficient less than or equal to 80(class 0), percent proficient greater than 80(class 1). There were 60 instances calculated in the machine learning algorithm. The accuracy was 48 percent, so it was no better than flipping a coin. This shows that the model was not fit for the data, and that the dataset has no correlation between the three key features.

Precision was the same as the accuracy because it was around 50 percent, completely random. However, the model was much better at predicting class 1. This is shown through the f1 score of 61 percent. The confusion matrix tells the same story as well. Scikit-learn was used for these models, and the confusion matrix shows a True Negative of 5, False Positive of 25, False Negative of 6, and a True Positive of 24. This matrix shows the complete lack of predictability and correlation of the 300 data instances. This is a great learning moment in that not all data is correlated or predictable, even if scaling and cleaning are used in an attempt to find a correlation.

– **Logistic Regression Results**

```
Accuracy: 0.48

Classification Report:
              precision    recall  f1-score   support

           0       0.45      0.17      0.24        30
           1       0.49      0.80      0.61        30

    accuracy                           0.48        60
   macro avg       0.47      0.48      0.43        60
weighted avg       0.47      0.48      0.43        60

Confusion Matrix:
[[ 5 25]
 [ 6 24]]
```

Fig. 17. This figure shows the precision, recall, f1-score, amount of instances and confusion matrix of the Logistic Regression Model.

5.4 Machine Learning Conclusion

– **Knowledge Gained from Machine Learning:**

The two features that were analyzed with machine learning models were expenditure per pupil and enrollment category. From the EDA, these two seemed to have the most interesting data stories to tell. However, it was found that there is no correlation between the two features and the target of reading "percent proficient". Even though this can feel like a defeat in a Data analyst's eyes, it really isn't. This just proves there is no correlation between those features, and that is knowledge gained from the machine learning portion of this report.

6 Conclusion

6.1 Closing Remarks

– **Visual of No Correlation in Tableau:**

Throughout the report, the question was asked if expenditures per pupil/district enrollment affected the reading proficiency of the 2017 Junior class. After thorough research and analysis, there is no correlation between the two features and reading proficiency. In other words, the two independent features are not valuable enough to determine reading proficiency. This is because student success is dependent on much more than how much the school spends and how big the school is. Future analysis should include features like county demographics, average county income, and reduced-lunch percentage. A final visual is shown below to show the lack of correlation between the 3 important attributes, and to reinforce this report's inability to find correlations between expense, district size, and reading proficiency.

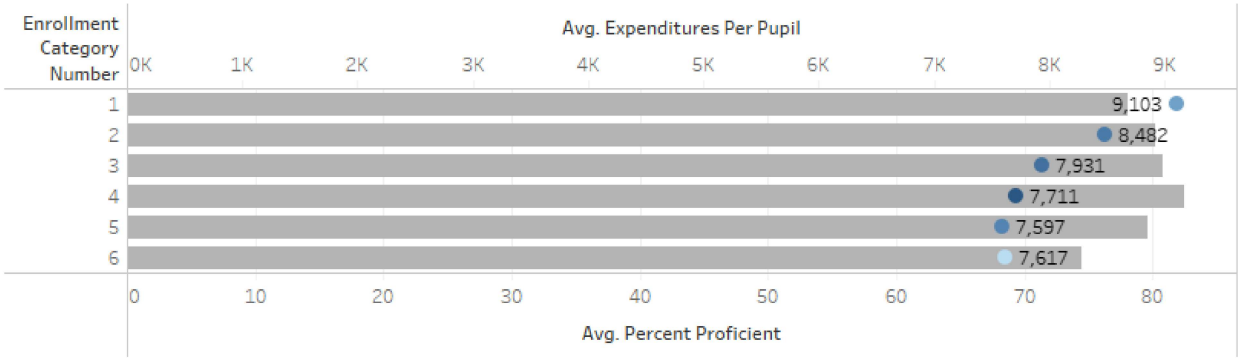


Fig. 18. This figure shows the lack of correlation between the 2 key features and the target of Reading Proficiency Percentage.

References

1. Dhaliwal, T.K., Bruno, P.: The rural/nonrural divide? K–12 district spending and implications of equity-based school funding. aera Open **7**, 2332858420982549 (2021), DOI: 10.1177/2332858420982549
2. Iowa Department of Education: Iowa school district expenditures by fiscal year (2024), https://data.iowa.gov/School-Finance/Iowa-School-District-Expenditures-by-Fiscal-Year/uutu-bzs3/about_data
3. Iowa Department of Education: Math and reading proficiency in iowa by school year (2024), https://data.iowa.gov/Primary-Secondary-Ed/Math-And-Reading-Proficiency-in-Iowa-by-School-Yea/f3h8-mnxi/about_data
4. Johnson, J.: More doesn't mean better: Larger high schools and more courses do not boost student achievement in iowa high schools. Rural School and Community Trust (2006), ERIC Document: ED491173
5. Science, T.D.: Top 10 algorithms for machine learning beginners. https://towardsdatascience.com/top-10-algorithms-for-machine-learning-beginners-149374935f3c, accessed: 2024-07-23