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. The writer will complete the Abstract last.

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 to invest their resources into their pupils. This project is going to show which Iowa Public Schools invested the most/least in their students. This report will also give the reader an idea of how much of an investment a school should spend to make their student the most successful they can possibly be, and whether there is a correlation between proficiency and investment per pupil. Success will be measured by proficiency scores in this report. These scores will be cleaned in PostgreSQL, analyzed through machine learning in Python and visualized 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, the writer will not be able to analyze every budget within the Iowa Public School system. The writer is going to keep it simple and analyze the budget every Iowa Public school used in the 2017 academic year: general instruction.

1.1 Goals of this Project

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

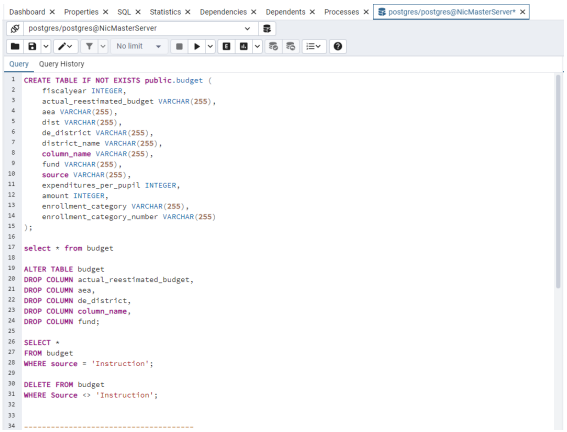
and copy statements for both CSV files. Second, the unnecessary attributes needed to be dropped from the tables to ensure simplicity and accuracy. Along with the unnecessary attributes, this report is only going to analyze reading proficiency. This decision was made because of the tight timeline of the 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**

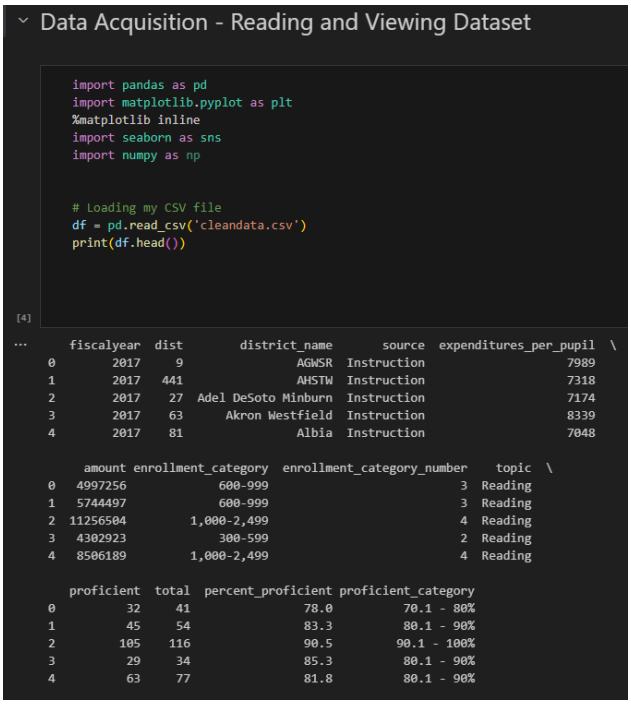


Fig. 3. Data Acquisition Process

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**

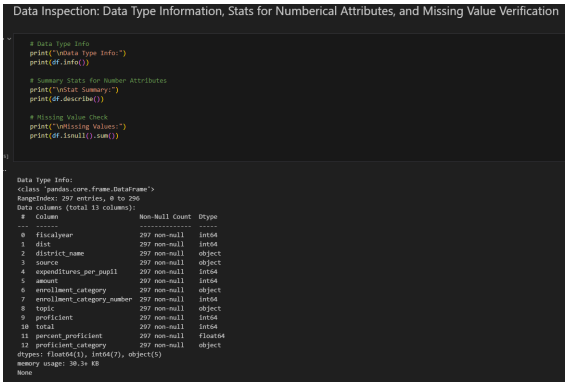


Fig. 4. Data Type information

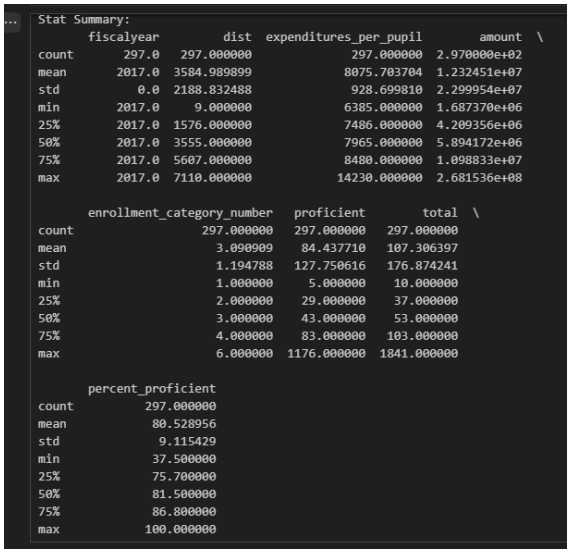


Fig. 5. Pandas Stats

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**

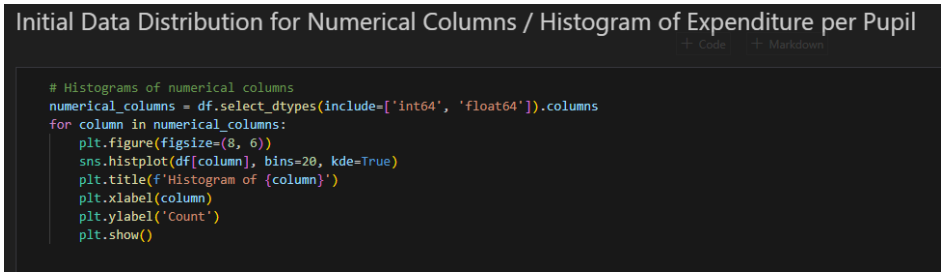


Fig. 6. Histogram Code

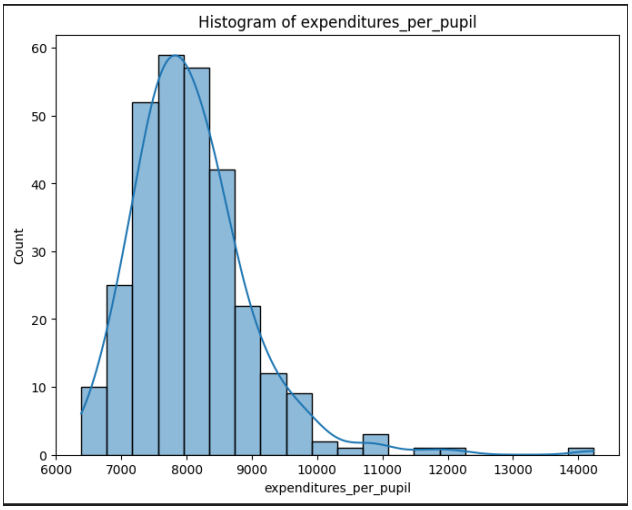


Fig. 7. Expenditure per Pupil Histogram showing Bell Curve around 8,000

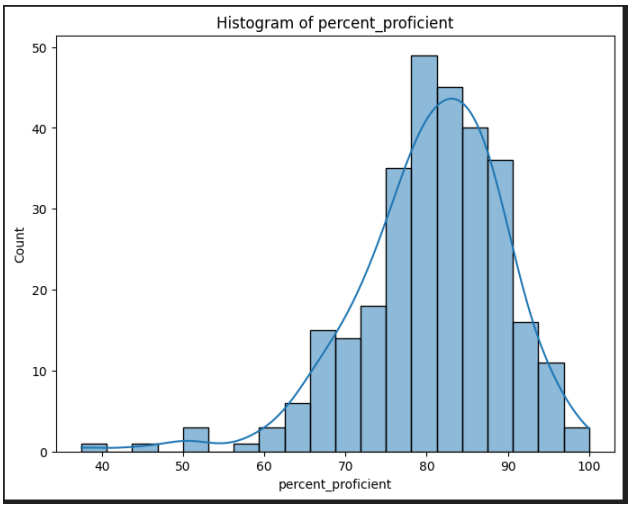


Fig. 8. Reading Proficiency Percentage Histogram showing Bell Curve around 80 percent

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**

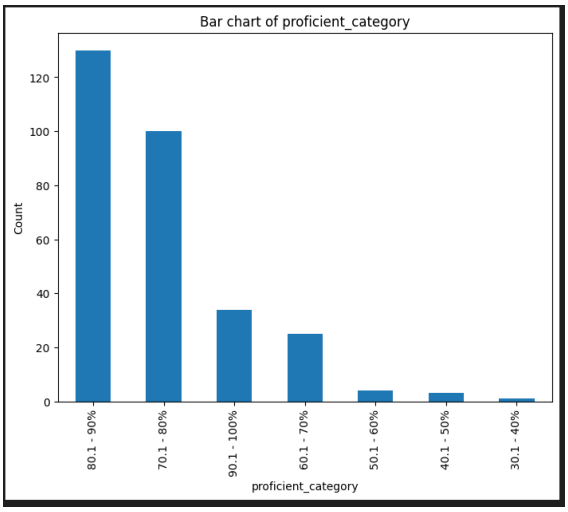


Fig. 9. Proficient Category Bar - most in 70-100 percent

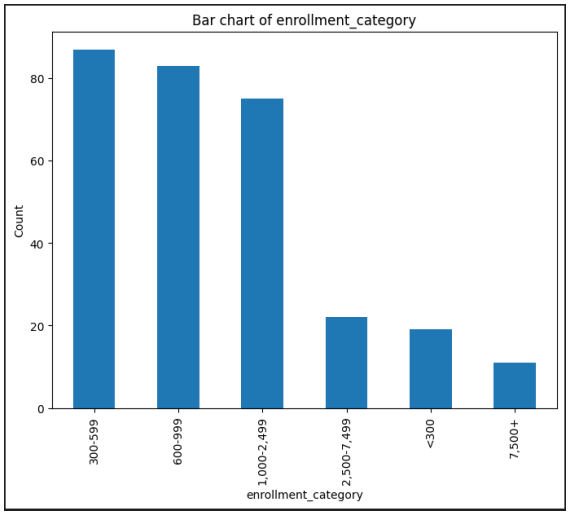


Fig. 10. Enrollment Category Bar - larger amount of smaller districts

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

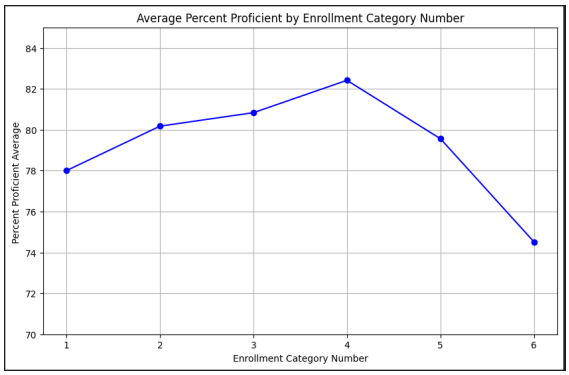


Fig. 12. Average Proficiency per District Size - district size 3-4 have highest reading proficiency

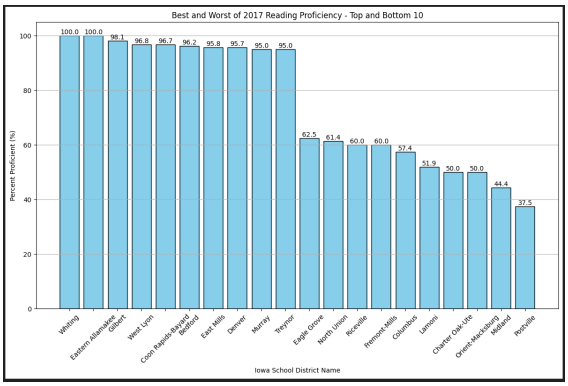


Fig. 13. Top/Bottom 10 in Reading Proficiency - the highs and lows of Iowa public education

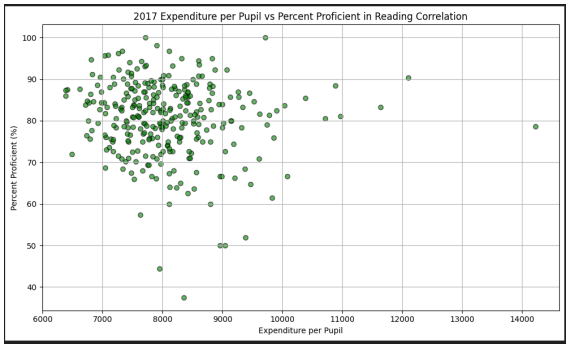


Fig. 14. Expense vs. Proficiency Correlation

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

6 Tableau Visualization of Results

7 Conclusion

□

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. 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