EXPLORING ACADEMIC PERFORMANCE TRENDS USING SAS ANOVA AND REGRESSION

1. Overview

Client:

A state-level education department in the United States focused on K–12 school performance evaluation

Objective:

To analyze academic performance across public schools using SAS and identify key demographic and institutional factors that explain differences in test scores. The goal was to inform policy changes and resource allocation.

2. Background

Standardized test results varied widely across schools, and the client lacked clarity on whether performance differences stemmed from school type, district funding, or student demographics. A robust statistical framework using SAS was needed to dissect these influences and provide actionable insights.

3. Data Summary

Dataset:

Academic records of 24,000 students across 120 public schools

Key Variables:

Variable	Type	Description
Math_Score	Continuous	Standardized math test score (0–100 scale) – dependent
School_Type	Categorical	Public / Charter / Magnet
Free_Lunch_Eligibility	Binary	Proxy for low-income status
Teacher_Student_Ratio	Continuous	Number of students per teacher
Parental_Education_Level	Ordinal	Less than High School / High School / College+

District_Funding_Per_Student	Continuous	Annual funding per student (USD)
Gender	Categorical	Male / Female

4. Methodology

Software Used:

SAS 9.4 (including PROC GLM, PROC REG, PROC MEANS, and PROC ANOVA)

SAS Workflow:

1. **Data Preprocessing:**
    - Encoded ordinal and categorical variables using PROC FORMAT
    - Cleaned and validated the dataset with PROC UNIVARIATE
    - Imputed missing values with PROC STDIZE (median strategy)
2. **Exploratory Analysis:**
    - Descriptive stats via PROC MEANS by school type
    - Distribution plots using PROC SGPLOT
    - Crosstabs with PROC FREQ for categorical factors
3. **ANOVA:**
    - PROC GLM used to analyze score differences across school types:
    - proc glm data=student_scores;
    - class School_Type;
    - model Math_Score = School_Type;
    - means School_Type / hovtest=levene;
    - run;
4. **Multiple Regression:**
    - PROC REG model including all predictors
    - Stepwise selection for optimal model
    - Checked for multicollinearity (VIF) and residual normality

5. Key Results

Predictor	Coefficient	p-value	Interpretation
Free_Lunch_Eligibility	–6.8	<0.001	Students from low-income families scored lower on average
Teacher_Student_Ratio	–1.3	0.027	Fewer students per teacher linked to higher scores
Parental_Education_Level	+5.4	<0.001	Higher parental education strongly associated with better scores
School_Type (Charter)	+2.1	0.043	Slight performance advantage for charter schools
District_Funding_Per_Student	Not Significant	0.18	No clear effect once other variables controlled

Model Fit:

- $R^2 = 0.63$
- Residuals approximately normal
- VIF < 1.8 (no multicollinearity concerns)

6. Visual Outputs (SAS):

- Boxplot of math scores by school type
- Regression coefficient plot with 95% CI
- Histogram of residuals
- Line chart: mean scores vs. teacher-student ratio

7. Deliverables

- Annotated .sas scripts for preprocessing, ANOVA, and regression
- Full report (20 pages) containing:
    - Summary statistics and visualizations
    - ANOVA tables and interpretation

    - Regression model with diagnostics and implications
- Stakeholder deck (5 slides):
    - Factors with greatest impact
    - Recommended funding and staffing strategies
    - District-wise performance insights

8. Application & Outcome

- Report used in state-level education funding hearings
- Recommendations adopted in strategic staffing allocation
- Charter school policies revisited to address performance parity

9. Strategic Value Delivered

- Uncovered **underlying drivers of score disparity** beyond raw averages
- Offered **data-backed justification for class-size reduction funding**
- Delivered **SAS-based framework** for annual academic performance reviews