School Finance: Which Factors Effect Public School Performance?

Swati Dontamsetti

Graduate School of Education, Rutgers University

Applied Multivariate Methods

Lauren E. Nolan, AICP

May 4, 2022

Abstract

Education is meant to be the great equalizer that will allow all children, regardless of ethnicity, sexuality, gender expression, or familial economic status, to rise through the socioeconomic ranks to achieve a better future than generations before them. However, schools just reproduce the same inequalities that exist in the larger society. And as we live in a capitalist society, money plays a prominent role in access to and quality of resources that schools have to support their students' growth and close inequity gaps. I use the Taxpayers' Guide to Education Spending (2020) and the New Jersey School Performance Reports (2016-2017 & 2018-2019) from the NJ Department of Education website to compare each district in NJ's spending activities and their school rankings. The data allows me to see how school spending and school rankings line up and investigate if there are specific categories of spending (i.e. teacher salaries, school supplies, support staff costs, etc.) that play a statistically significant role in the rankings of schools in NJ. While a district's poverty level has the biggest impact on school rankings, I find three areas where increased spending significantly impacts school ranking.

Keywords: school finance, poverty, equity, per-pupil spending, school rankings

School Finance: Which Factors Effect Public School Performance?

Horace Mann, the pioneer of the system of American public schools that is used today, in the mid-19th century called education the "great equalizer of the conditions of men" (Duncan, 2018). The dream was that, regardless of ethnicity, racial identity, sexuality, gender, or familial economic status, all students in the United States would have the ability to attend free, quality public schools and gain the knowledge necessary to live the American Dream and become economically better off than generations before. But as Arne Duncan (2018), Secretary of Education under President Obama, points out, "Too often, the difference between a life of promise and a life in peril hinges not on a student's potential but on the quality of the local public school." To investigate where the best-performing public schools are spending their money, I look into the public schools of New Jersey.

New Jersey has consistently ranked in the top 3 best performing public school systems in the country. And, in the last two years, it has held the number 1 spot. Moreover, NJ has over 600 highly localized public schools. And the history of *Abbott* legislation in NJ has made school funding equity the state law (MacInnes, 2009). However, schools in NJ are highly segregated by ethnicity and socioeconomic status, which directly impacts the school's performance. As with elsewhere in the country, families with the means can move to districts with better-performing schools. In this way, education is not the "great equalizer," and poverty plays a significant role in a school's overall performance and ranking in the state. Despite the inequality present, NJ spends explicitly more on schools in high-poverty districts. I am interested in seeing which factors districts can spend their money on to improve their school's performance.

I use the public data from the Taxpayers' Guide to Education Spending (2020) and the New Jersey School Performance Reports (2016-2017 & 2018-2019) from the NJ Department of

Education website to identify important variables that affect school ranking. I look only at K12 districts in NJ, which drops the overall district country from over 600 to 220, so that I can plausibly track how the district's schools affect student graduation rates and academic performance from the student entering Kindergarten to them leaving after 12th grade. I conduct a multiple regression on three different models and ANOVA tests to identify the model that best fits the data. When poverty rates are held constant, I hypothesize that class supplies, support services, student-teacher ratio, and median support personnel salary have a statistically significant impact on school rankings.

Literature Review

In reviewing the system of education in the United States, Mickelson et al. (2018) note that "[e]ducational equity and excellence remain much sought after and elusive societal goals" (p. 128). In order to have true equity, students of different means and backgrounds would get the resources they need to perform on par with their wealthy and neurotypical peers. We know that demographics and concentrations of poverty do play a part in how well a school can educate its students. Mickelson et al. (2018) note, "[e]ven after controlling for student background and prior achievement, researchers find that achievement and attainment are worse for those who attended racially segregated [] schools" (p. 121). While there has not been much push to consolidate districts or to desegregate schools after *Parents Involved in Community Schools [PICS]* 2007 declared voluntary desegregation plans unconstitutional if they used students' race as a criterion for admission (Riel et al., 2018), NJ gone through numerous *Abbott* lawsuits that have continuously amended the state's funding model to explicitly provide more support to high-poverty districts (MacInnes, 2009). The question then remains, how do schools spend their money to most effectively improve students' academic achievements?

Research shows that "communication, learning facilities, proper guidance and family stress are the factors that affect the student performance. … Communication, learning facilities and proper guidance shows the positive impact on the student performance and the family stress shows the negative impact on the student performance but the significant level is high" (Mushtaq & Khan, 2012, p. 21). And of those factors, teacher quality (Lamas, 2015) and emotional supports (Wong, 2016) seem to be the most important. In Boston, for example, "the school system is homing in on how childhood trauma can undermine achievement and developing means for helping kids cope with it" (Wong, 2016).

The factors within a school's control are the quality of its teachers, the maintenance of its facilities, and the guidance it offers to its students. These are the three broad categories that I investigate further in my analysis of NJ public schools. I hypothesize, based on the research, that student-teacher ratio, years of teaching experience, percentage of per-pupil funding that goes to classroom supplies, and support staff personnel will be statistically significant determinants of a district's rankings, which could have policy implications for how districts can most successfully spend their funding to increase academic achievement for all their students.

Data

I use the Taxpayers' Guide to Education Spending (2020) and 2016-17 & 2018-19 New Jersey School Performance Reports from the NJ Department of Education to compare each district's spending activities against its rankings. I am only exploring the districts that serve Kindergarten through 12^{th} grade so that a student can spend their entire school career in the same district. Most of the data reported in the New Jersey School Performance Reports is based on data submitted by school districts through NJ SMART data collections. For assessments and exams scored through outside vendors, such as statewide assessments or SAT exams, the outside

vendor provides student performance data. The Taxpayers' Guide to Education Spending (2020) actual expenditure data for the 2018-19 school year originate from the district's Comprehensive Annual Financial Reports certified by the districts' public-school accountants. The district submits the enrollment and staffing data. My dependent variables are the ELA and Math NJSLA "met expectations" percentages for grade 10 (from 2016-17, when the 12th graders would have been in 10th grade) and 4-year High School graduation rates from the NJ School Performance Reports.

From the NJ School Performance Reports, my independent variables are the percentage of economically disadvantaged students (eligible for free and reduced lunch), Students with Disabilities (students who are classified for special education), and years of teaching experience. From the Taxpayers' Guide to Education Spending, my independent variables are Budgetary Per Pupil Cost, Classroom Salaries and Benefits, Classroom Supplies/Textbooks, Classroom Purchased Services/Other Costs, Total Support Services, Total Operations and Maintenance of Plant, Extracurricular Costs, Total Equipment Cost, Ratio of Students to Classroom Teachers and Median Classroom Teacher Salary, and Ratio of Students to Educational Support Personnel and Median Salary.

When considering how to rank schools, I turned to US News and Niche, both websites that consistently put out school rankings. Unfortunately, I could not download the rankings from the websites themselves in a way that matched the NJ district schools' data. Instead, based on their ranking system, I used the information I had at my disposal to create school rankings of my own. I am only looking at K12 districts; that way, it could be argued that one school district affected the entirety of a student's academic performance from Kindergarten to graduation. While both US News and Niche use numerous factors in creating school rankings, I only use

ELA and Math state test scores and the Graduation rate of each district. 10th grade, the last years students take state testing, ELA test scores are easy enough to find, but Math scores are a bit more complicated. Students take different math tests in 10th grade, depending on the academic track. Some students take Geometry in 10th grade and some students take Algebra 2 (Linden Public Schools, p. 44); therefore, I take an average of both scores to measure Math performance. This is an imperfect system, as there might be younger and older students taking each test, but it is the closest approximation of 10th grade Math state test performance that I could achieve. I then create an overall state performance value based on the average ELA and Math scores. Both US News and Niche give academic performance more weight than a school's graduation rate. In creating my ranking, I give twice the weight to overall state performance as to graduation rates.

Although I do not have the data on students who stay in one district their entire school tenure, by measuring only K12 districts, I could reasonably extrapolate based on the data how a district's spending would affect the graduation rate and academic performance of a student who did spend their entire schooling in that district.

Methods

After cleaning and combining my different datasets based on district ID, I am left with 176 districts to analyze. I run three different multiple regressions on increasingly more specific models. To do this, I must first check that my data meet the assumptions of a regression equation. The base model for my multiple regression is the following:

$$H_o\colon \beta_{PerPup\$} = \beta_{StuTeachRat} = \beta_{MedTeacSal} = \beta_{AvgTeachExpYr} = \beta_{\%FRL} = 0$$

$$H_a\colon \beta_{PerPup\$} \neq \beta_{StuTeachRat} \neq \beta_{MedTeacSal} \neq \beta_{AvgTeachExpYr} \neq \beta_{\%FRL} \neq 0$$

I am looking to analyze how significant the amount of money per pupil, the student-teacher ratio, the median teacher salary, average years of teaching experience, and the percent of

students on free-and-reduced lunch in each district are towards that district's school ranking. In models 2 and 3, I look further into specific factors of per-pupil spending and various special education teachers and support staff that also make up a school. My full model, in comparison, has 17 dependent variables.

Figure 1 shows that my independent variable, School Rank, has slight peaks above and below the center, but the data is normally distributed without any outliers. Moreover, while Graduation Rate (Figure 2) does have numerous outliers, the state test scores (Figure 3) do not, and with more weight given to the test scores than the graduation rate in School Rank, this seems to normalize.

As this is a social sciences analysis, the correlation between most of my variables could be categorized as weak. However, when using the correlation scale in education circles of 0.2-0.4 equaling moderate correlation and anything over 0.4 as a strong correlation, I have a few moderately correlated variables and one that is strongly correlated. The percentage of students on free-and-reduced lunch seems to be strongly negatively correlated with school rank, but even more so on graduation rates and test scores (which make up my school rank variable).

To investigate this further, I create a factor variable that categorizes each district's poverty level from very low to very high. Figure 8 shows that my base model variables have a considerable range in districts regardless of the poverty level. So, very high poverty and very low poverty districts are spending varying amounts per pupil; they have varying student-teacher ratios, and the average years of experience of the teachers in their district vary, as does the median salary of their teachers. We can conclude that a district's poverty level is not correlated with any of the other independent variables. But Figure 9 shows the high, negative linear correlation between a district's poverty level and its high school graduation rate and test scores.

Since I am concerned mainly with how *well-performing* schools spend their money, including the level of district poverty as an independent variable would allow me to hold it constant in my analysis of other factors that affect school rankings.

Homoscedasticity is present, and there is no autocorrelation among the variables I use in my models. Having passed all the assumptions of a linear regression, I run my three models. The results are shown in Figures 4, 5, and 6.

After conducting an ANOVA test to measure the goodness-of-fit, shown in Figure 7, I find that model 2 is a better fit than model 1 (my base model). But between model 2 (all variables not included in model 1) and model 3 (the full model), either could be the better fit. If I use an alpha=0.1, then model 3 is the better fit, but if I use an alpha=0.05, then model 2 is the better fit. Moreover, model 2 has an adjusted-R^2 of 0.61, and model 3 has an adjusted-R^2 of 0.62. Both models account for about 60% of the variance in the dependent variable that can be explained by the independent variables. As both models are similar and provide valuable insights, I use the results of both model 2 and model 3 in making my conclusions.

Results

Model 2

$$H_o\colon \beta_{\%ClassSal} = \beta_{\%ClassSup} = \beta_{\%ClassServ} = \beta_{\%SupServ} = \beta_{\%Ops} = \beta_{\%Extra} = \beta_{\%Equip} = \beta_{Enrollment} = \beta_{\%FRL} = \beta_{\%Disabled} = \beta_{\%ELL} = \beta_{StuSupPerRat} = \beta_{MedSupPerSal} = 0$$

$$H_a\colon \beta_{\%ClassSal} \neq \beta_{\%ClassSup} \neq \beta_{\%ClassServ} \neq \beta_{\%SupServ} \neq \beta_{\%Ops} \neq \beta_{\%Extra} \neq \beta_{\%Equip} \neq \beta_{Enrollment} \neq \beta_{\%FRL} \neq \beta_{\%Disabled} \neq \beta_{\%ELL} \neq \beta_{StuSupPerRat} \neq \beta_{MedSupPerSal} \neq 0$$

The p-value of the model is very small, less than 0.001. Therefore, we can reject the null hypothesis in favor of the alternate hypothesis. Certain independent variables are statistically significant predictors of the dependent variable (School Rank). With an adjusted-R^2 of 0.61, the model accounts for 61% of the variance in the dependent variable that can be explained by the independent variables. Looking closely at the independent variables, at an alpha=0.05, aside from the demographic data of the student body (percent on free-and-reduced lunch, percent disabled, and percent classified as ELL students), only Median Support Personnel Salary shows as statistically significant. But when I lower the alpha level to 0.1, then the percent of per-pupil spending towards Support Services and Class Supplies is also statistically significant.

Based on the results, we can say that a $1 increase in Median Teacher Salary is associated with a 0.00014620-point increase in School Rank when holding all other variables constant. This does not seem to have much of an impact on School Rank; however, when holding all other variables constant, a 1 percentage point increase in per-pupil spending towards Support Services is associated with a 64.3-point increase in School Rank. And, when holding all other variables constant, a 1 percentage point increase in per-pupil spending that goes towards Class Supplies is associated with a 164.7-point in School Rank.

Model 3

$$H_o:\ \beta_{PerPup\$} = \beta_{StuTeachRat} = \beta_{MedTeacSal} = \beta_{AvgTeachExpYr} = \beta_{\%ClassSal} = \beta_{\%ClassSup} = \beta_{\%ClassServ} = \beta_{\%SupServ} = \beta_{\%Ops} = \beta_{\%Extra} = \beta_{\%Equip} = \beta_{Enrollment} = \beta_{\%FRL} = \beta_{\%Disabled} = \beta_{\%ELL} = \beta_{StuSupPerRat} = \beta_{MedSupPerSal} = 0$$

$$H_a:\ \beta_{PerPup\$} \neq \beta_{StuTeachRat} \neq \beta_{MedTeacSal} \neq \beta_{AvgTeachExpYr} \neq \beta_{\%ClassSal} \neq \beta_{\%ClassSup} \neq \beta_{\%ClassServ} \neq \beta_{\%SupServ} \neq \beta_{\%Ops} \neq \beta_{\%Extra} \neq \beta_{\%Equip} \neq \beta_{Enrollment} \neq \beta_{\%FRL} \neq \beta_{\%Disabled} \neq \beta_{\%ELL} \neq \beta_{StuSupPerRat} \neq \beta_{MedSupPerSal} \neq 0$$

The p-value of the model is very small, less than 0.001. Therefore, we can reject the null hypothesis in favor of the alternate hypothesis. Certain independent variables are statistically significant predictors of the dependent variable (School Rank). With an adjusted-R^2 of 0.62, the model accounts for 62% of the variance in the dependent variable that can be explained by the independent variables. Looking closely at the independent variables, at an alpha=0.05, aside from the demographic data of the student body (percent on free-and-reduced lunch and percent disabled), only Student-Teacher Ratio shows as statistically significant. But when I lower the alpha level to 0.1, then Per Pupil spending is also statistically significant.

Based on the results, we can say that when holding all other variables constant, a 1-point increase in the student-teacher ratio (meaning more students for every teacher) is associated with a 1.69-point increase in the School Rank. And, when holding all other variables constant, a $1 increase in per-pupil spending is associated with a 0.00102141 increase in School Rank. Neither variable seems to affect School Rank all that much.

Student Demographics

In comparison, student demographic data is the most statistically significant variable. In both Model 2 and 3, the percent of students on free-and-reduced lunch, disabled, and classified as ELL have a similar coefficient value and p-values, so we look specifically at the effect of the variables in Model 3. When holding all other variables constant, a 1 percentage point increase in ELL is associated with a 0.31961411 increase in School Rank. And, when holding all other variables constant, a 1 percentage point increase in Disabled students is associated with a 0.54106294 decrease in School Rank. And, when holding all other variables constant, a 1 percentage point increase in students on free-and-reduced lunch is associated with a 0.41254739

decrease in School Rank. Despite student demographic data being more statistically significant, it does not change the School Rank too much.

Discussion

Limitations and Room for Further Study

It is important to know that School Rank is a value from zero to one hundred. Therefore, in Model 2, when a 1 percentage point increase in per-pupil spending towards Class Supplies was associated with a 164.7-point in School Rank, it brings attention to the critical limitations of the data. A plateauing effect of all these variables on School Rank might exist that a linear regression does not account for. The student-teacher ratio is another area different from what is usually discussed – the model's results note that an increase in the number of students per teacher positively affects School Rank. This might be associated with the fact that special education classrooms have smaller class sizes, and therefore districts with high special education students would have a smaller student-teacher ratio. And, because special education students are ordinarily exempt from state testing, the nuances of this variable are challenging to tease apart. Moreover, my School Rank marker is very simplistic; it does not account for all the variables that US News and Niche consider – perhaps, the most important being the academic performance of minority students in the district.

In further studying the data, I would also be interested in seeing if years of Principal or even Superintendent experience, as the school leadership, affects the school's rank. Moreover, it might be prudent to specifically look into the high-poverty school districts to analyze their academic performance and how they spend their money. If I were to expand the analysis provided here, I would have accounted for schools that hold the historic *Abbott* designation, as

some are still highly-poor while some have since gentrified, to see how their per-pupil spending breaks down.

Policy Implication

The models' results can classify the statistically significant variables into student demographics, staffing, and per-pupil spending. The variables most statistically significant are the student demographics. Schools with higher levels of poverty and students who are disabled negatively effects School Rank, while higher levels of students who are ELL positively affect School Rank. Still, the overall point value effect on School Rank is not even 1-point. But when we hold student demographics constant, money spent on class supplies and support services, the district's student-teacher ratio, and median support personnel salary have statistically significant effects on School Rank.

The NJ Department of Education classifies class supplies as all the materials needed for all the district's classes, such as calculators, microscopes, textbooks, tablets, laptops, workbooks, tests, markers, paper, pencils, paints, and other classroom supplies are included. Support personnel are the counselors, librarians, nurses, child study team members, and other educational support services personnel in the district. And support services consist of:

> Attendance, social work, health and guidance services, educational media/school library services and child study team services. … This area also includes the costs associated with physical and mental health services that are not direct instruction, but are nevertheless provided to students, such as supervision of health services, health appraisal (including screening for vision, communicable diseases, and hearing deficiencies), screening for psychiatric services, periodic health examinations, emergency injury and illness care, dental services, nursing services, and communications with parents and

> medical officials. The expenditures of the guidance office includes counseling, record maintenance, and placement services. The costs for the child study team include salaries and benefits for members related to the development and evaluation of student individualized education programs (IEPs). … School library services include books repairs, audiovisual services, educational television services, and computer assisted instruction services. (State of New Jersey Department of Education, p. 7)

Taken together, support services and support personnel are the services and staff members who supplement the teaching process by assessing and improving students' well-being. And model 2's results show that the percentage of per-pupil spending that goes towards class supplies and support services has the most significant positive effect on School Rank (164.7 and 64.3, respectively).

From the analysis results, we can conclude that school districts that spend more of their money on classroom supplies, whole-child support services, and retaining high-quality support personnel will be able to increase the academic achievement and graduation rate (the components of school rank) of their students. As a state, we need to encourage districts to focus more on taking care of the whole child, not just their classroom performance, but their social and emotional well-being are equally important. The literature supports this finding; when districts care for their students' mental and emotional health, they perform better academically.

References

Duncan, A. (2018, May 25). *Education: The "Great Equalizer."* Encyclopedia Britannica. https://www.britannica.com/topic/Education-The-Great-Equalizer-2119678

Lamas, H. A. (2015). School Performance. *Propósitos y Representaciones*, *3*(1), 351–386.

Linden Public Schools. (n.d.). *High School Course Description Guide 2022-2023*. https://2aeea1fd-c428-4021-bff3-cea43f1fe971.filesusr.com/ugd/1f806c_3afcc63f109b4e1684c80f755d7abe80.pdf

Lloyd, S. C., & Harwin, A. (2020, September 1). *Nation's Schools Get a 'C' Once Again, Even as Pandemic Turns Up the Heat*. Education Week. https://www.edweek.org/policy-politics/nations-schools-get-a-c-once-again-even-as-pandemic-turns-up-the-heat/2020/09

MacInnes, G. (2009). *In Plain Sight: Simple, Difficult Lessons from New Jersey's Expensive Efforts to Close the Achievement Gap*. Century Foundation Press.

Mickelson, R. A., Giersch, J., Nelson, A. H., & Bottia, M. C. (2018). Do Charter Schools Undermine Efforts to Create Racially and Socioeconomically Diverse Public Schools? In *Choosing Charters: Better Schools or More Segregation?* (pp. 116–132). Teachers College Press.

Morse, R., & Brooks, E. (2022, April 25). *How US News Calculated the 2022 Best High Schools Rankings*. USNews.com. https://www.usnews.com/education/best-high-schools/articles/how-us-news-calculated-the-rankings

Mushtaq, I., & Khan, S. N. (2012). Factors Affecting Students' Academic Performance. *Global Journal of Management and Business Research*, *12*(9), 17–22.

Niche.com Inc. (n.d.). *The Best School Districts Methodology*. Niche. https://www.niche.com/about/methodology/best-school-districts/

Riel, V., Parcel, T. L., Mickelson, R. A., & Smith, S. S. (2018). Do magnet and charter schools exacerbate or ameliorate inequality? *Sociology Compass*, *12*(9).

State of New Jersey. (n.d.). *2016-2017 New Jersey School Performance Reports: Reference Guide*. NJ School Performance Report. https://navilp7rg08njprsharedst.blob.core.windows.net/perf-reports-ct/Documents/1617/ReferenceGuide.pdf

State of New Jersey. (n.d.). *2018-2019 New Jersey School Performance Reports: Reference Guide*. NJ School Performance Report. https://navilp7rg08njprsharedst.blob.core.windows.net/perf-reports-ct/Documents/1819/ReferenceGuide.pdf

State of New Jersey Department of Education. (n.d.). *Taxpayers' Guide to Education Spending – 2020*. School Finance. https://www.nj.gov/education/guide/2020/INTROTGES.pdf

Wong, A. (2016, May 23). *What Are Massachusetts Public Schools Doing Right?* The Atlantic. https://www.theatlantic.com/education/archive/2016/05/what-are-massachusetts-public-schools-doing-right/483935/

Appendix A - Figures

Figure 1

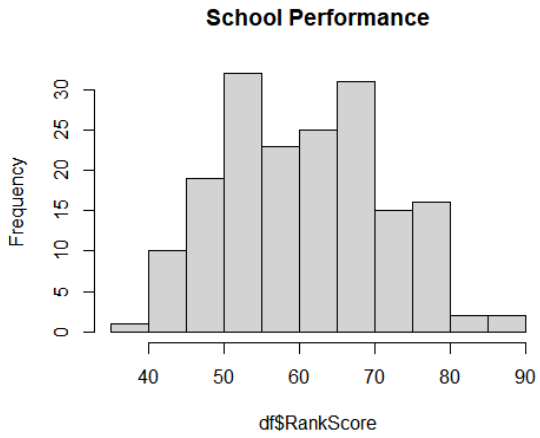


Figure 2

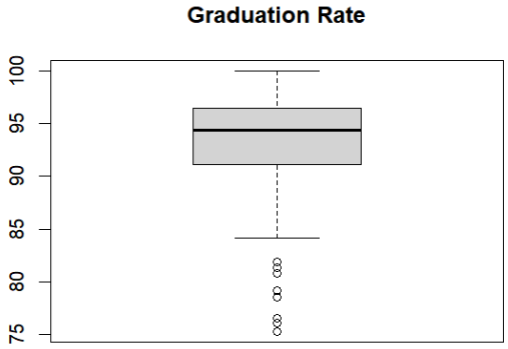


Figure 3

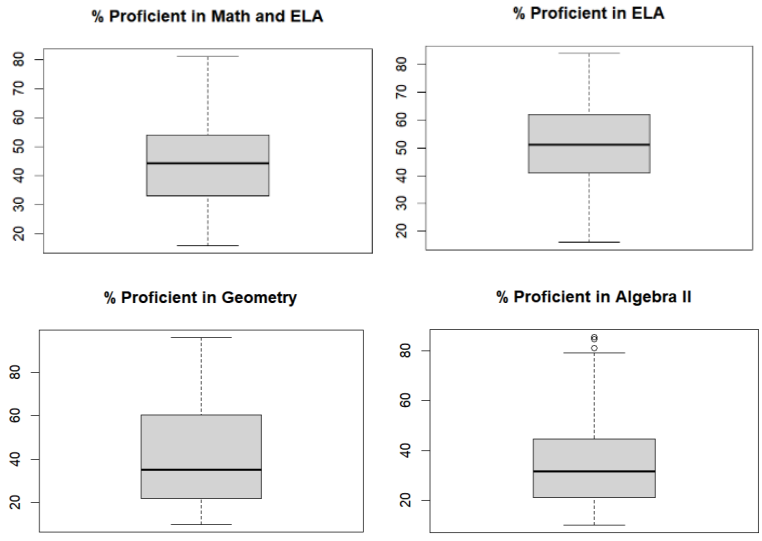


Figure 4

```
Call:
lm(formula = df$RankScore ~ df$PerPupTot + df$StuTeachRat + df$MedTeachSal +
    df$TeachAvgYearsExp + df$PercentFRL)

Residuals:
    Min      1Q  Median      3Q     Max
-18.392  -4.273  -0.329   4.576  17.274

Coefficients:
                     Estimate  Std. Error t value             Pr(>|t|)
(Intercept)         17.34560042 12.48656076   1.389             0.166607
df$PerPupTot         0.00114975  0.00041681   2.758             0.006443 **
df$StuTeachRat       2.52444919  0.73827932   3.419             0.000786 ***
df$MedTeachSal       0.00018903  0.00007947   2.378             0.018494 *
df$TeachAvgYearsExp -0.67253073  0.34166092  -1.968             0.050646 .
df$PercentFRL       -0.35172251  0.02429063 -14.480 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.875 on 170 degrees of freedom
Multiple R-squared:  0.5918,    Adjusted R-squared:  0.5798
F-statistic: 49.29 on 5 and 170 DF,  p-value: < 0.00000000000000022
```

Figure 5

```
Call:
lm(formula = df$RankScore ~ df$PercentTotClassSal + df$PercentTotClassSup +
    df$PercentTotClassServ + df$PercentTotSupServ + df$PercentTotOps +
    df$PercentTotExtra + df$PercentTotEquip + df$EnrollmentTot +
    df$PercentFRL + df$PercentDisabled + df$PercentELL + df$StuSupPerRat +
    df$MedSupPerSal)

Residuals:
     Min       1Q   Median       3Q      Max
-13.0454  -4.4766  -0.0056   4.1865  17.4357

Coefficients:
                         Estimate   Std. Error t value             Pr(>|t|)
(Intercept)            42.13493843  29.92923751   1.408              0.16110
df$PercentTotClassSal  19.10295982  35.06459455   0.545              0.58664
df$PercentTotClassSup 164.69346030  88.09891912   1.869              0.06337 .
df$PercentTotClassServ 48.55458229  49.94489919   0.972              0.33242
df$PercentTotSupServ   64.25837939  37.01388621   1.736              0.08445 .
df$PercentTotOps        3.05388956  44.66679455   0.068              0.94558
df$PercentTotExtra     -7.91347968  81.07799211  -0.098              0.92237
df$PercentTotEquip    102.34156101  89.73120057   1.141              0.25575
df$EnrollmentTot        0.00004072   0.00014322   0.284              0.77654
df$PercentFRL          -0.43871129   0.03766455 -11.648 < 0.0000000000000002 ***
df$PercentDisabled     -0.50613899   0.18633094  -2.716              0.00732 **
df$PercentELL           0.41672252   0.16203604   2.572              0.01102 *
df$StuSupPerRat         0.00461099   0.04183351   0.110              0.91237
df$MedSupPerSal         0.00014620   0.00005416   2.700              0.00768 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.593 on 162 degrees of freedom
Multiple R-squared:  0.6423,    Adjusted R-squared:  0.6136
F-statistic: 22.37 on 13 and 162 DF,  p-value: < 0.00000000000000022
```

Figure 6

```
Call:
lm(formula = df$RankScore ~ df$PerPupTot + df$PercentTotClassSal +
    df$PercentTotClassSup + df$PercentTotClassServ + df$PercentTotSupServ +
    df$PercentTotOps + df$PercentTotExtra + df$PercentTotEquip +
    df$EnrollmentTot + df$PercentFRL + df$PercentDisabled + df$PercentELL +
    df$StuTeachRat + df$MedTeachSal + df$StuSupPerRat + df$MedSupPerSal +
    df$TeachAvgYearsExp)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8437  -4.6913   0.0438   3.7475  16.6861

Coefficients:
                          Estimate   Std. Error t value             Pr(>|t|)
(Intercept)           18.89190094  34.71310837   0.544              0.58705
df$PerPupTot           0.00102141   0.00055098   1.854              0.06563 .
df$PercentTotClassSal  13.54031863  37.97067874   0.357              0.72187
df$PercentTotClassSup 142.74745642  88.02445378   1.622              0.10687
df$PercentTotClassServ 38.41535553  51.93193050   0.740              0.46056
df$PercentTotSupServ   51.77586666  37.92840208   1.365              0.17416
df$PercentTotOps      -30.32062724  48.30441254  -0.628              0.53111
df$PercentTotExtra     -7.41699153  82.11504712  -0.090              0.92814
df$PercentTotEquip    105.96069126  89.12468596   1.189              0.23626
df$EnrollmentTot        0.00006002   0.00014361   0.418              0.67657
df$PercentFRL          -0.41254739   0.03926051 -10.508 < 0.0000000000000002 ***
df$PercentDisabled     -0.54106294   0.18817096  -2.875              0.00459 **
df$PercentELL           0.31961411   0.17969338   1.779              0.07722 .
df$StuTeachRat          1.69283553   0.81717399   2.072              0.03993 *
df$MedTeachSal          0.00009673   0.00010549   0.917              0.36052
df$StuSupPerRat         0.01403805   0.05368400   0.261              0.79405
df$MedSupPerSal         0.00002999   0.00007451   0.402              0.68791
df$TeachAvgYearsExp    -0.14100751   0.35485979  -0.397              0.69164
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.507 on 158 degrees of freedom
Multiple R-squared:  0.6602,    Adjusted R-squared:  0.6236
F-statistic: 18.06 on 17 and 158 DF,  p-value: < 0.00000000000000022
```

Figure 7

```
> anova(rank_simple, rank_part, test='F') #partial is better
Analysis of Variance Table

Model 1: df$RankScore ~ df$PerPupTot + df$StuTeachRat + df$MedTeachSal +
    df$TeachAvgYearsExp + df$PercentFRL
Model 2: df$RankScore ~ df$PercentTotClassSal + df$PercentTotClassSup +
    df$PercentTotClassServ + df$PercentTotSupServ + df$PercentTotOps +
    df$PercentTotExtra + df$PercentTotEquip + df$EnrollmentTot +
    df$PercentFRL + df$PercentDisabled + df$PercentELL + df$StuSupPerRat +
    df$MedSupPerSal
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    170 8036.2
2    162 7042.2  8    994.07 2.8585 0.005338 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(rank_part, rank_all, test='F') #all only slightly better
Analysis of Variance Table

Model 1: df$RankScore ~ df$PercentTotClassSal + df$PercentTotClassSup +
    df$PercentTotClassServ + df$PercentTotSupServ + df$PercentTotOps +
    df$PercentTotExtra + df$PercentTotEquip + df$EnrollmentTot +
    df$PercentFRL + df$PercentDisabled + df$PercentELL + df$StuSupPerRat +
    df$MedSupPerSal
Model 2: df$RankScore ~ df$PerPupTot + df$PercentTotClassSal + df$PercentTotClassSup +
    df$PercentTotClassServ + df$PercentTotSupServ + df$PercentTotOps +
    df$PercentTotExtra + df$PercentTotEquip + df$EnrollmentTot +
    df$PercentFRL + df$PercentDisabled + df$PercentELL + df$StuTeachRat +
    df$MedTeachSal + df$StuSupPerRat + df$MedSupPerSal + df$TeachAvgYearsExp
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    162 7042.2
2    158 6689.8  4    352.35 2.0804 0.08586 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8

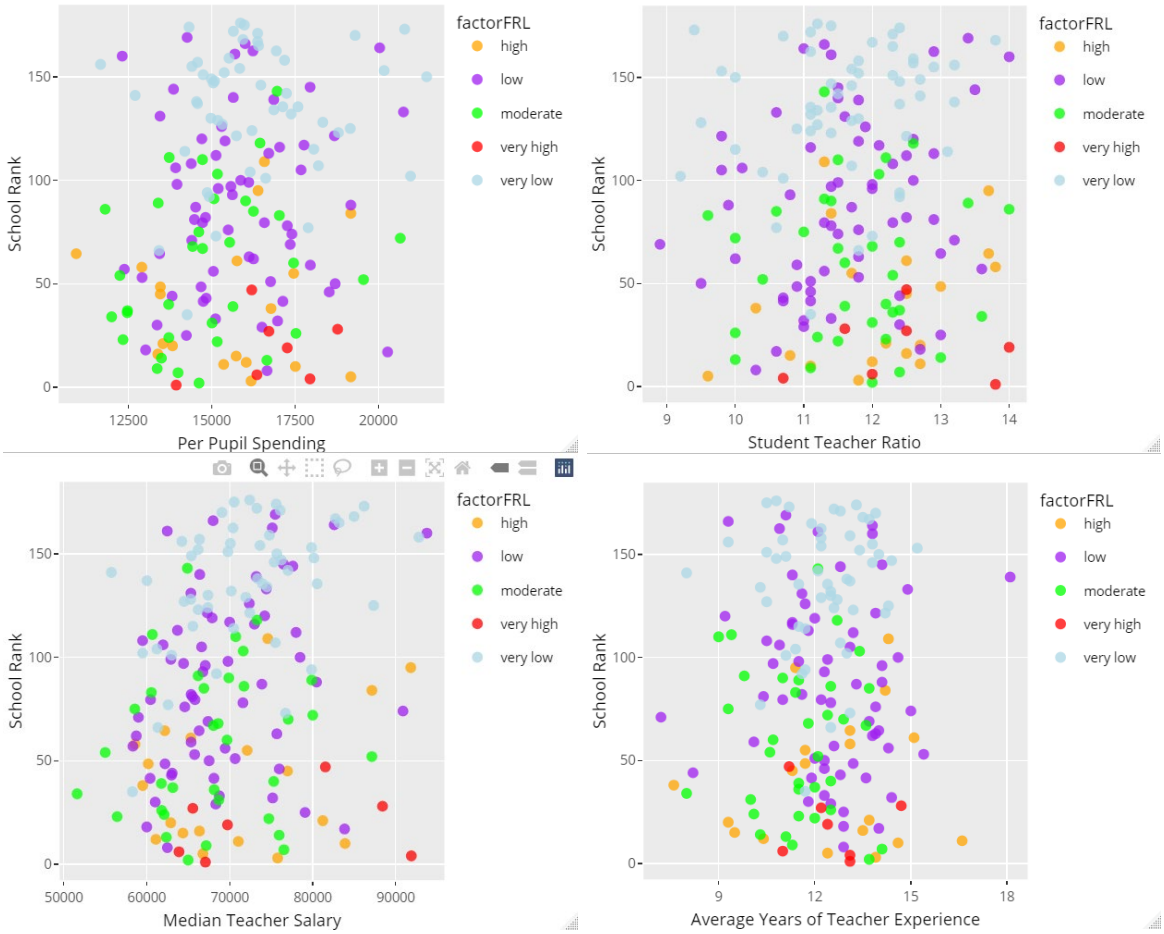
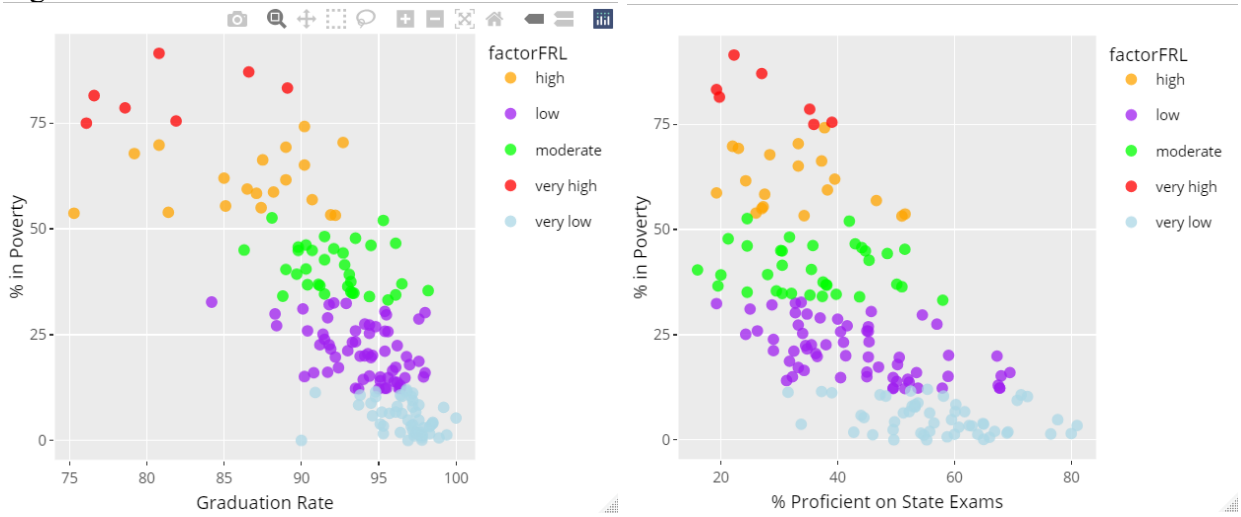


Figure 9



Appendix B – R Code

```
#install packages
install.packages("tidyverse")
install.packages("dplyr")
install.packages("lmtest")
install.packages("ggplot2")
install.packages("ggpubr")
install.packages("car")
install.packages("plotly")
install.packages("plyr")

#load libraries
library(tidyverse)
library(dplyr)
library(lmtest)
library(ggplot2)
library(ggpubr)
library(car)
library(plotly)
library(plyr)

#set working directory
setwd("C:\\Users\\swati\\OneDrive - Rutgers University\\2 Spring 2022\\App Multivar
Methods\\Final Project")

#LOAD DATA
budget_ppcost_df <- read.csv("CSG1.csv")
class_salaries_df <- read.csv("CSG3.csv")
class_supplies_df <- read.csv("CSG4.csv")
class_services_df <- read.csv("CSG5.csv")
tot_support_df <- read.csv("CSG6.csv")
ops_maintain_df <- read.csv("CSG10.csv")
extracurricular_df <- read.csv("CSG13.csv")
tot_equipment_df <- read.csv("CSG15.csv")
StuTeachRatio_df <- read.csv("CSG16.csv")
StuSpecRatio_df <- read.csv("CSG17.csv")
enrollment_df <- read.csv("EnrollmentTrendsByStudentGroup.csv")
teachers_exp_df <- read.csv("TeachersExperience.csv")
gradratetrends_df <- read.csv("GraduatonRateTrendsProgress.csv")
dropoutrate_df <- read.csv("DropoutRateTrends.csv")
math_scores <- read.csv("MATHperformance16-17.csv")
ela_score <- read.csv("ELAperformance16-17.csv")

#remove scientific notation
options(scipen=999)
```

```
#only keep k12 districts
test <- budget_ppcost_df[!(budget_ppcost_df$GROUP=="A. K-6" |
        budget_ppcost_df$GROUP=="B. K-8 / 0 - 400" | budget_ppcost_df$GROUP=="C. K-8 /
        401 - 750" | budget_ppcost_df$GROUP=="D. K-8 / 751 +" |
        budget_ppcost_df$GROUP=="H. 7-12 / 9-12" | budget_ppcost_df$GROUP=="I. CSSD"
        | budget_ppcost_df$GROUP=="J. VOC" | budget_ppcost_df$GROUP=="K. Charter" |
        budget_ppcost_df$CONAME=="Statewide"),]
budget_ppcost <- subset(test, select = -c(PP11, RK11, RK21, PP31, RK31, E11, E31, GROUP))
budget_ppcost$DIST <- as.numeric(budget_ppcost$DIST)

test1 <- class_salaries_df[!(class_salaries_df$GROUP=="A. K-6" |
        class_salaries_df$GROUP=="B. K-8/0-400" | class_salaries_df$GROUP=="C. K-8/401-
        750" | class_salaries_df$GROUP=="D. K-8/751+" | class_salaries_df$GROUP=="H. 7-
        12/9-12" | class_salaries_df$GROUP=="I. CSSD" | class_salaries_df$GROUP=="J.
        Voc" | class_salaries_df$GROUP=="K. Charter" |
        class_salaries_df$CONAME=="Statewide"),]
class_salaries <- subset(test1, select = -c(PP13, RK13, PCT13, SBA3, RK23, PP33, RK33,
        PCT33, SBC3, GROUP))
class_salaries$DIST <- as.numeric(class_salaries$DIST)

test2 <- class_services_df[-c(1:291, 524:705),]
test2 <- test2[!(test2$CONAME=="Statewide"),]
class_services <- subset(test2, select = -c(PP15, RK15, PCT15, RK25, PP35, RK35, PCT35,
        GROUP))
class_services$DIST <- as.numeric(class_services$DIST)

test3 <- class_supplies_df[-c(1:291, 524:705),]
test3 <- test3[!(test3$CONAME=="Statewide"),]
class_supplies <- subset(test3, select = -c(PP14, RK14, PCT14, RK24, PP34, RK34, PCT34,
        GROUP))
class_supplies$DIST <- as.numeric(class_supplies$DIST)

test4 <- tot_support_df[-c(1:291, 524:705),]
test4 <- test4[!(test4$CONAME=="Statewide"),]
tot_support <- subset(test4, select = -c(PP16,RK16,PCT16,RK26,PP36,RK36,PCT36,GROUP))
tot_support$DIST <- as.numeric(tot_support$DIST)

test5 <- ops_maintain_df[-c(1:291, 524:705),]
test5 <- test5[!(test5$CONAME=="Statewide"),]
ops_maintain <- subset(test5, select = -c(PP110, RK110, PCT110, RK210, PP310, RK310,
        PCT310, GROUP))
ops_maintain$DIST <- as.numeric(ops_maintain$DIST)

test6 <- extracurricular_df[-c(1:291, 524:705),]
test6 <- test6[!(test6$CONAME=="Statewide"),]
```

```
extracurricular <- subset(test6, select = -c(PP113, RK113, PCT113, RK213, PP313, RK313,
        PCT313, GROUP))
extracurricular$DIST <- as.numeric(extracurricular$DIST)

test7 <- tot_equipment_df[-c(1:291, 524:705),]
test7 <- test7[!(test7$CONAME=="Statewide"),]
tot_equipment <- subset(test7, select = -c(PP115,PP315,GROUP))
tot_equipment$DIST <- as.numeric(tot_equipment$DIST)

test8 <- StuTeachRatio_df[-c(1:291, 524:705),]
test8 <- test8[!(test8$CONAME=="Statewide"),]
StuTeachRatio <- subset(test8, select = -c(RK0016, RKSAL0016, STRAT0116, RK0116,
        SALT0116, RKSAL0116, GROUP))
StuTeachRatio$DIST <- as.numeric(StuTeachRatio$DIST)

test9 <- StuSpecRatio_df[-c(1:291, 524:705),]
test9 <- test9[!(test9$CONAME=="Statewide"),]
StuSpecRatio <- subset(test9, select = -c(RK0017, RKSAL0017, SSRAT0117, RK0117,
        SALS0117, RKSAL0117, GROUP))
StuSpecRatio$DIST <- as.numeric(StuSpecRatio$DIST)

enrollment_data <- enrollment_df[,-c(1:2, 4:6, 10:13)]
enrollment_data$DistrictCode <- as.numeric(enrollment_data$DistrictCode)

teachers_exp <- teachers_exp_df[,-c(1:2, 4:6, 8:12)]
teachers_exp$DistrictCode <- as.numeric(teachers_exp$DistrictCode)

test12 <- gradratetrends_df[,-c(1:2, 4, 8:10)]
test12 <- test12[!(test12$CohortYear == 2016),]
test12 <- test12[!(test12$CohortYear == 2017),]
test12 <- test12[!(test12$CohortYear == 2018),]
grad_rate <- test12
grad_rate$DistrictCode <- as.numeric(grad_rate$DistrictCode)

dropout_rate <- dropoutrate_df[,-c(1:2, 4, 6)]
dropout_rate$DistrictCode <- as.numeric(dropout_rate$DistrictCode)

#isolate and then merge test scores
ELA_grad10 <- ela_score[(ela_score$Grade_Subject == "Grade 10"),]
ELA_grad10$MetExcExpPerc <- as.numeric(ELA_grad10$MetExcExpPerc)
ELA_grad10$DistrictCode <- as.numeric(ELA_grad10$DistrictCode)
ELA_grad10 <- ELA_grad10[complete.cases(ELA_grad10),]
ELA_grad10.m <- ddply(ELA_grad10, .(DistrictCode), summarize,
        MetExcExpPerc.ELA=mean(MetExcExpPerc))

MATH_grad10 <- math_scores[!(math_scores$Grade_Subject == "Grade 3"),]
```

```
MATH_grad10 <- MATH_grad10[!(MATH_grad10$Grade_Subject == "Grade 4"),]
MATH_grad10 <- MATH_grad10[!(MATH_grad10$Grade_Subject == "Grade 5"),]
MATH_grad10 <- MATH_grad10[!(MATH_grad10$Grade_Subject == "Grade 6"),]
MATH_grad10 <- MATH_grad10[!(MATH_grad10$Grade_Subject == "Grade 7"),]
MATH_grad10 <- MATH_grad10[!(MATH_grad10$Grade_Subject == "Grade 8"),]
MATH_grad10 <- MATH_grad10[!(MATH_grad10$Grade_Subject == "Algebra I"),]
MATH_grad10$MetExcExpPerc <- as.numeric(MATH_grad10$MetExcExpPerc)
MATH_grad10 <- MATH_grad10[complete.cases(MATH_grad10),]
MATH_algebraII <- MATH_grad10[(MATH_grad10$Grade_Subject == "Algebra II"),]
MATH_algebraII.m <- ddply(MATH_algebraII, .(DistrictCode), summarize,
        MetExcExpPerc.AlgII=mean(MetExcExpPerc))
MATH_geometry <- MATH_grad10[(MATH_grad10$Grade_Subject == "Geometry"),]
MATH_geometry.m <- ddply(MATH_geometry, .(DistrictCode), summarize,
        MetExcExpPerc.Geo=mean(MetExcExpPerc))

math_combo <- merge(MATH_algebraII.m, MATH_geometry.m, by.x = "DistrictCode")
math_combo$MetExcExpPerc.MC <- (math_combo$MetExcExpPerc.AlgII +
        math_combo$MetExcExpPerc.Geo)/2

testscores_combo <- merge(math_combo, ELA_grad10.m, by.x="DistrictCode")
testscores_combo$MetExcExpPerc.Avg <- (testscores_combo$MetExcExpPerc.MC +
        testscores_combo$MetExcExpPerc.ELA)/2

#merge Taxpayers' Guide to Education Spending (TGES) dataframes
m <- merge(budget_ppcost, class_salaries, by.x = c("DIST", "CONAME", "DISTNAME"), by.y
        = c("DIST", "CONAME", "DISTNAME"))
m1 <- merge(m, class_supplies, by.x = c("DIST", "CONAME", "DISTNAME"), by.y =
        c("DIST", "CONAME", "DISTNAME"))
m2 <- merge(m1, class_services, by.x = c("DIST", "CONAME", "DISTNAME"), by.y =
        c("DIST", "CONAME", "DISTNAME"))
m3 <- merge(m2, tot_support, by.x = c("DIST", "CONAME", "DISTNAME"), by.y = c("DIST",
        "CONAME", "DISTNAME"))
m4 <- merge(m3, ops_maintain, by.x = c("DIST", "CONAME", "DISTNAME"), by.y =
        c("DIST", "CONAME", "DISTNAME"))
m5 <- merge(m4, extracurricular, by.x = c("DIST", "CONAME", "DISTNAME"), by.y =
        c("DIST", "CONAME", "DISTNAME"))
m6 <- merge(m5, tot_equipment, by.x = c("DIST", "CONAME", "DISTNAME"), by.y =
        c("DIST", "CONAME", "DISTNAME"))
m6$PP215 <- as.numeric(m6$PP215)
m6$PCT215 <- round(m6$PP215/m6$PP21, digits = 3)
m7 <- merge(m6, StuTeachRatio, by.x = c("DIST", "CONAME", "DISTNAME"), by.y =
        c("DIST", "CONAME", "DISTNAME"))
tges_df <- merge(m7, StuSpecRatio, by.x = c("DIST", "CONAME", "DISTNAME"), by.y =
        c("DIST", "CONAME", "DISTNAME"))
```

```
#merge New Jersey School Performance Reports (SPR) dataframes
m8 <- merge(enrollment_data, teachers_exp, by.x = "DistrictCode")
m9 <- merge(m8, grad_rate, by.x = "DistrictCode")
m10 <- merge(m9, testscores_combo, by.x = "DistrictCode")
spr_df <- merge(m10, dropout_rate, by.x = "DistrictCode")

#merge both datasets for final dataframe and create csv
data_df <- merge(tges_df, spr_df, by.x = "DIST", by.y = "DistrictCode")

#change column names
colnames(data_df)[colnames(data_df) == "DIST"] <- "DistCode"
colnames(data_df)[colnames(data_df) == "CONAME"] <- "County"
colnames(data_df)[colnames(data_df) == "DISTNAME"] <- "District"
colnames(data_df)[colnames(data_df) == "PP21"] <- "PerPupTot"
colnames(data_df)[colnames(data_df) == "E21"] <- "EnrollmentTot"
colnames(data_df)[colnames(data_df) == "PP23"] <- "PerPupClassSal"
colnames(data_df)[colnames(data_df) == "PCT23"] <- "PercentTotClassSal"
colnames(data_df)[colnames(data_df) == "PP24"] <- "PerPupClassSup"
colnames(data_df)[colnames(data_df) == "PCT24"] <- "PercentTotClassSup"
colnames(data_df)[colnames(data_df) == "PP25"] <- "PerPupClassServ"
colnames(data_df)[colnames(data_df) == "PCT25"] <- "PercentTotClassServ"
colnames(data_df)[colnames(data_df) == "PP26"] <- "PerPupSupServ"
colnames(data_df)[colnames(data_df) == "PCT26"] <- "PercentTotSupServ"
colnames(data_df)[colnames(data_df) == "PP210"] <- "PerPupOps"
colnames(data_df)[colnames(data_df) == "PCT210"] <- "PercentTotOps"
colnames(data_df)[colnames(data_df) == "PP213"] <- "PerPupExtra"
colnames(data_df)[colnames(data_df) == "PCT213"] <- "PercentTotExtra"
colnames(data_df)[colnames(data_df) == "PP215"] <- "PerPupEquip"
colnames(data_df)[colnames(data_df) == "PCT215"] <- "PercentTotEquip"
colnames(data_df)[colnames(data_df) == "STRAT0016"] <- "StuTeachRat"
colnames(data_df)[colnames(data_df) == "SALT0016"] <- "MedTeachSal"
colnames(data_df)[colnames(data_df) == "SSRAT0017"] <- "StuSupPerRat"
colnames(data_df)[colnames(data_df) == "SALS0017"] <- "MedSupPerSal"
colnames(data_df)[colnames(data_df) == "Economically.Disadvantaged.Students"] <-
        "PercentFRL"
colnames(data_df)[colnames(data_df) == "Students.with.Disabilities"] <- "PercentDisabled"
colnames(data_df)[colnames(data_df) == "English.Learners"] <- "PercentELL"
colnames(data_df)[colnames(data_df) == "TeacherAvgYearsExp_District"] <-
        "TeachAvgYearsExp"
colnames(data_df)[colnames(data_df) == "Dropout_District"] <- "DropoutRate"

final_df <- subset(data_df, select = -c(SBB3, CohortYear, GraduationRateType, PerPupClassSal,
        PerPupClassSup, PerPupClassServ, PerPupSupServ, PerPupOps, PerPupExtra,
        PerPupEquip))
```

```
#check datatypes
summary(final_df)

#set variables as numeric
final_df$PercentTotClassSup <- as.numeric(final_df$PercentTotClassSup)
final_df$PercentTotClassServ <- as.numeric(final_df$PercentTotClassServ)
final_df$PercentTotSupServ <- as.numeric(final_df$PercentTotSupServ)
final_df$PercentTotOps <- as.numeric(final_df$PercentTotOps)
final_df$PercentTotExtra <- as.numeric(final_df$PercentTotExtra)
final_df$StuTeachRat <- as.numeric(final_df$StuTeachRat)
final_df$MedTeachSal <- as.numeric(final_df$MedTeachSal)
final_df$StuSupPerRat <- as.numeric(final_df$StuSupPerRat)
final_df$MedSupPerSal <- as.numeric(final_df$MedSupPerSal)
final_df$PercentFRL <- as.numeric(final_df$PercentFRL)
final_df$PercentDisabled <- as.numeric(final_df$PercentDisabled)
final_df$PercentELL <- as.numeric(final_df$PercentELL)
final_df$TeachAvgYearsExp <- as.numeric(final_df$TeachAvgYearsExp)
final_df$GraduationRate <- as.numeric(final_df$GraduationRate)
final_df$DropoutRate <- as.numeric(final_df$DropoutRate)

#drop NAs
final_noNA <- final_df[complete.cases(final_df),]

#write CSV
write.csv(final_noNA, "C:\\Users\\swati\\OneDrive - Rutgers University\\2 Spring 2022\\App
        Multivar Methods\\Final Project\\finalprojectdata.csv", row.names = FALSE)

#LOAD DATA
df <- read.csv("finalprojectdata.csv")

#create School Rank variable
df$RankScore <- ((df$MetExcExpPerc.Avg*2/3)+(df$GraduationRate/3))
df$Rank <- rank(df$RankScore)

#check for linear regression assumptions
#1 variables are submitted by districts, shouldn't contain errors
hist(df$MetExcExpPerc.AlgII) #right skewed
hist(df$MetExcExpPerc.Geo) #right skewed
hist(df$MetExcExpPerc.MC) #very slight right skew
hist(df$MetExcExpPerc.ELA) #normal distribution
hist(df$MetExcExpPerc.Avg) #relatively normal
hist(df$GraduationRate) #left skewed, seems to have lots of outliers
hist(df$RankScore, main = "School Performance") #normal, use this as dependent

boxplot(df$MetExcExpPerc.AlgII) #a few top performing outliers
boxplot(df$MetExcExpPerc.Geo) #no outliers
```

```
boxplot(df$MetExcExpPerc.MC) #no outliers
boxplot(df$MetExcExpPerc.ELA) #no outliers
boxplot(df$MetExcExpPerc.Avg) #no outliers
boxplot(df$GraduationRate) #seems to have lots of outliers
boxplot(df$RankScore) #no outliers, use this as dependent

#2 model is correctly specified:
rank_all <-
        lm(df$RankScore~df$PerPupTot+df$PercentTotClassSal+df$PercentTotClassSup+df$Pe
        rcentTotClassServ+df$PercentTotSupServ+df$PercentTotOps+df$PercentTotExtra+df$P
        ercentTotEquip+df$EnrollmentTot+df$PercentFRL+df$PercentDisabled+df$PercentELL
        +df$StuTeachRat+df$MedTeachSal+df$StuSupPerRat+df$MedSupPerSal+df$TeachAvg
        YearsExp)
summary(rank_all) #adjusted r2 = 0.6236
bptest(rank_all)

rank_part <-
        lm(df$RankScore~df$PercentTotClassSal+df$PercentTotClassSup+df$PercentTotClassS
        erv+df$PercentTotSupServ+df$PercentTotOps+df$PercentTotExtra+df$PercentTotEquip
        +df$EnrollmentTot+df$PercentFRL+df$PercentDisabled+df$PercentELL+df$StuSupPer
        Rat+df$MedSupPerSal)
summary(rank_part) #adjusted r2 = 0.6136
bptest(rank_part)

rank_simple <-
        lm(df$RankScore~df$PerPupTot+df$StuTeachRat+df$MedTeachSal+df$TeachAvgYear
        sExp+df$PercentFRL)
summary(rank_simple) #adjusted r2 = 0.5798
bptest(rank_simple)

#3 check for linear relationship: < 0.1 negligible ;0.1-0.2 weak; 0.2-0.4 moderate; > 0.4 strong
plot(df$PerPupTot, df$MetExcExpPerc.Avg)
cor(df$PerPupTot, df$MetExcExpPerc.Avg) #0.169 weak postive relationship

plot(df$StuTeachRat, df$MetExcExpPerc.Avg)
cor(df$StuTeachRat, df$MetExcExpPerc.Avg) #0.035 no relationship

plot(df$TeachAvgYearsExp, df$MetExcExpPerc.Avg)
cor(df$TeachAvgYearsExp, df$MetExcExpPerc.Avg) #0.003 no relationship

plot(df$PercentFRL, df$MetExcExpPerc.Avg)
cor(df$PercentFRL, df$MetExcExpPerc.Avg) #-0.656 strong negative relationship

plot(df$GraduationRate, df$MetExcExpPerc.Avg)
cor(df$GraduationRate, df$MetExcExpPerc.Avg) #0.526 strong positive relationship
```

```
plot(df$PerPupTot, df$GraduationRate)
cor(df$PerPupTot, df$GraduationRate) #0.072 no relationship

plot(df$StuTeachRat, df$GraduationRate)
cor(df$StuTeachRat, df$GraduationRate) #-0.124 weak negative relationship

plot(df$TeachAvgYearsExp, df$GraduationRate)
cor(df$TeachAvgYearsExp, df$GraduationRate) #-0.087 no relationship

plot(df$PercentFRL, df$GraduationRate)
cor(df$PercentFRL, df$GraduationRate) #-0.785 strong negative relationship

plot(df$PerPupTot, df$RankScore)
cor(df$PerPupTot, df$RankScore) #0.165 weak positive relationship

plot(df$StuTeachRat, df$RankScore)
cor(df$StuTeachRat, df$RankScore) #0.014 no relationship

plot(df$TeachAvgYearsExp, df$RankScore)
cor(df$TeachAvgYearsExp, df$RankScore) #-0.01 no relationship

plot(df$PercentFRL, df$RankScore)
cor(df$PercentFRL, df$RankScore) #-0.716 strong negative relationship

#4 zero mean
#5 normality of error term - 176 observations no need to check because of CLT
df$resids <- residuals(rank_all)
hist(df$resids) #dependent = rank seems to be more normally distributed

#6 errors are homoscedastic
df$predvals <- fitted(rank_all)
plot(df$predvals, df$resids)

bptest(rank_all)

#7 no autocorrelation
plot(df$PerPupTot, df$resids)
cor(df$PerPupTot, df$resids)

plot(df$PercentTotClassSal, df$resids)
cor(df$PercentTotClassSal, df$resids)

plot(df$PercentTotClassSup, df$resids)
cor(df$PercentTotClassSup, df$resids)

plot(df$PercentTotClassServ, df$resids)
```

```
cor(df$PercentTotClassServ, df$resids)

plot(df$PercentTotSupServ, df$resids)
cor(df$PercentTotSupServ, df$resids)

plot(df$PercentTotOps, df$resids)
cor(df$PercentTotOps, df$resids)

plot(df$PercentTotExtra, df$resids)
cor(df$PercentTotExtra, df$resids)

plot(df$PercentTotEquip, df$resids)
cor(df$PercentTotEquip, df$resids)

plot(df$EnrollmentTot, df$resids)
cor(df$EnrollmentTot, df$resids)

plot(df$PercentFRL, df$resids)
cor(df$PercentFRL, df$resids)

plot(df$PercentDisabled, df$resids)
cor(df$PercentDisabled, df$resids)

plot(df$PercentELL, df$resids)
cor(df$PercentELL, df$resids)

plot(df$StuTeachRat, df$resids)
cor(df$StuTeachRat, df$resids)

plot(df$MedTeachSal, df$resids)
cor(df$MedTeachSal, df$resids)

plot(df$StuSupPerRat, df$resids)
cor(df$StuSupPerRat, df$resids)

plot(df$MedSupPerSal, df$resids)
cor(df$MedSupPerSal, df$resids)

plot(df$TeachAvgYearsExp, df$resids)
cor(df$TeachAvgYearsExp, df$resids)

#8 no multicollinearity
var_only <- df[,c(4:29)]
cor(var_only) #gradrate:frl = -0.782; gradrate:dropout = -0.825; math:ela = 0.874; math|ela:avg >
        0.96
```

```
#goodness of fit test
anova(rank_simple, rank_part, test='F') #partial is better
anova(rank_part, rank_all, test='F') #all only slightly better

summary(rank_all)

#create graphs of perpuptot against graduation rate, change size for poverty
for (i in 1:176){
  if(df$PercentFRL[i] < 12.20){
    df$factorFRL[i] = "very low"
  } else if(df$PercentFRL[i] < 32.78) {
    df$factorFRL[i] = "low"
  } else if(df$PercentFRL[i] < 52.80) {
    df$factorFRL[i] = "moderate"
  } else if(df$PercentFRL[i] < 75) {
    df$factorFRL[i] = "high"
  } else {
    df$factorFRL[i] = "very high"
  }
}

plot1 <- df %>%
  ggplot(aes(StuTeachRat,Rank, color=factorFRL,
             text = paste("School:", District,
                          "<br>%FRL:", PercentFRL))) +
  geom_point(alpha=0.75, size=2) +
  scale_color_manual(values = c("very low" = "light blue",
                                "low" = "purple",
                                "moderate" = "green",
                                "high" = "orange",
                                "very high" = "red")) +
  labs(y="School Rank", x="Student Teacher Ratio")
ggplotly(plot1, tooltip = "text")

plot2 <- df %>%
  ggplot(aes(PerPupTot,Rank, color=factorFRL,
             text = paste("School:", District,
                          "<br>%FRL:", PercentFRL))) +
  geom_point(alpha=0.75, size=2) +
  scale_color_manual(values = c("very low" = "light blue",
                                "low" = "purple",
                                "moderate" = "green",
                                "high" = "orange",
                                "very high" = "red")) +
  labs(y="School Rank", x="Per Pupil Spending")
ggplotly(plot2, tooltip = "text")
```

```
plot3 <- df %>%
  ggplot(aes(MedTeachSal,Rank, color=factorFRL,
             text = paste("School:", District,
                          "<br>%FRL:", PercentFRL))) +
  geom_point(alpha=0.75, size=2) +
  scale_color_manual(values = c("very low" = "light blue",
                                "low" = "purple",
                                "moderate" = "green",
                                "high" = "orange",
                                "very high" = "red")) +
  labs(y="School Rank", x="Median Teacher Salary")
ggplotly(plot3, tooltip = "text")

plot4 <- df %>%
  ggplot(aes(TeachAvgYearsExp,Rank, color=factorFRL,
             text = paste("School:", District,
                          "<br>%FRL:", PercentFRL))) +
  geom_point(alpha=0.75, size=2) +
  scale_color_manual(values = c("very low" = "light blue",
                                "low" = "purple",
                                "moderate" = "green",
                                "high" = "orange",
                                "very high" = "red")) +
  labs(y="School Rank", x="Average Years of Teacher Experience")
ggplotly(plot4, tooltip = "text")

plot5 <- df %>%
  ggplot(aes(GraduationRate, PercentFRL, color=factorFRL,
             text = paste("School:", District,
                          "<br>%FRL:", PercentFRL))) +
  geom_point(alpha=0.75, size=2) +
  scale_color_manual(values = c("very low" = "light blue",
                                "low" = "purple",
                                "moderate" = "green",
                                "high" = "orange",
                                "very high" = "red")) +
  labs(x="Graduation Rate", y="% in Poverty")
ggplotly(plot5, tooltip = "text")

plot6 <- df %>%
  ggplot(aes(MetExcExpPerc.Avg, PercentFRL, color=factorFRL,
             text = paste("School:", District,
                          "<br>%FRL:", PercentFRL))) +
  geom_point(alpha=0.75, size=2) +
  scale_color_manual(values = c("very low" = "light blue",
```

```
                        "low" = "purple",
                        "moderate" = "green",
                        "high" = "orange",
                        "very high" = "red")) +
  labs(x="% Proficient on State Exams", y="% in Poverty")
ggplotly(plot6, tooltip = "text")

#some graphs for presentation
summary(df$PerPupTot)
boxplot(df$PerPupTot, main="Per Pupil Spending")

summary(df$MetExcExpPerc.Avg)
boxplot(df$MetExcExpPerc.Avg, main="% Proficient in Math and ELA")
summary(df$MetExcExpPerc.ELA)
boxplot(df$MetExcExpPerc.ELA, main="% Proficient in ELA")
summary(df$MetExcExpPerc.AlgII)
boxplot(df$MetExcExpPerc.AlgII, main="% Proficient in Algebra II")
summary(df$MetExcExpPerc.Geo)
boxplot(df$MetExcExpPerc.Geo, main="% Proficient in Geometry")

summary(df$GraduationRate)
boxplot(df$GraduationRate, main="Graduation Rate")
```