Predictors of Academic Performance in Ontario Elementary Schools

Sami Shahid

2025-04-04

Setting Up

We first import the necessary libraries and read the Excel data. Then, we clean the dataset by removing irrelevant columns, converting percentage strings to numeric values.

1. Description of the Dataset

```
##  [1] "Board Name"
##  [2] "School Name"
##  [3] "School Type"
##  [4] "School Level"
##  [5] "School Language"
##  [6] "Grade Range"
##  [7] "City"
##  [8] "Enrolment"
##  [9] "Latitude"
## [10] "Longitude"
## [11] "Percentage of Students Whose First Language Is Not English"
## [12] "Percentage of Students Whose First Language Is Not French"
## [13] "Percentage of Students Who Are New to Canada from a Non-English Speaking Country"
## [14] "Percentage of Students Who Are New to Canada from a Non-French Speaking Country"
## [15] "Percentage of Students Receiving Special Education Services"
## [16] "Percentage of Students Identified as Gifted"
## [17] "Percentage of Grade 3 Students Achieving the Provincial Standard in Reading"
## [18] "Percentage of Grade 3 Students Achieving the Provincial Standard in Writing"
## [19] "Percentage of Grade 3 Students Achieving the Provincial Standard in Mathematics"
## [20] "Percentage of Grade 6 Students Achieving the Provincial Standard in Reading"
## [21] "Percentage of Grade 6 Students Achieving the Provincial Standard in Writing"
## [22] "Percentage of Grade 6 Students Achieving the Provincial Standard in Mathematics"
## [23] "Percentage of School-Aged Children Who Live in Low-Income Households"
## [24] "Percentage of Students Whose Parents Have No Degree, Diploma or Certificate"
```

- **Board Name, School Name, School Type, School Level, School Language, Grade Range**: Identifiers for the school and board, including type and grade levels offered (6 variables)
- **City, Latitude, Longitude**: Location information, including the city and geographic coordinates of the school (3 variables)
- **Enrolment**: Number of students enrolled at the school (1 variable)

- **Percentage of Students Whose First Language Is Not English, Percentage of Students Whose First Language Is Not French, Percentage of Students Who Are New to Canada from a Non-English Speaking Country, Percentage of Students Who Are New to Canada from a Non-French Speaking Country**: Demographic information related to students' language background and immigrant status (4 variables)
- **Percentage of Students Receiving Special Education Services, Percentage of Students Identified as Gifted**: Proportions of students receiving special education services or identified as gifted (2 variables)
- **Percentage of Grade 3 and 6 Students Achieving the Provincial Standard in Reading, Writing, Mathematics**: Academic achievement rates for students in various grades across different subjects (6 variables)
- **Percentage of School-Aged Children Who Live in Low-Income Households, Percentage of Students Whose Parents Have No Degree, Diploma or Certificate** : Socioeconomic data on student household income and parental education levels (2 variables)

2. Background of the Datset

This dataset was provided by the Ministry of Education of Ontario and contains information on publicly funded schools (JK-12) and student demographics across the province. It includes EQAO assessment results from the 2022-2023 academic year for Grades 3, 6, and 9, as well as OSSLT results. Also, it provides demographic data for each school based on the 2021-2022 school year. The dataset was collected to monitor and improve the quality of publicly funded education in Ontario and is used as a key resource for analyzing trends in school performance.

3. Overall Research Question

Goal: To understand how various factors (such as School Type, Enrollment, Socioeconomic Factors, etc.) influence academic performance and school demographics across elementary school's in Ontario

Specific Questions:

1. How does the academic performance vary between different school types and language?
2. How does the percentage of low-income households correlate with academic performance?
3. How does the percentage of parents with degree correlate with academic performance?
4. What are the top 3 highest and lowest-performing school boards by overall score?
5. How does geographical location impact academic performance?
6. Do schools with higher enrolment perform better or worse?
7. What factors have a significant and insignificant impact on academic performance?

4. Tables

4.1 Average Grade 3 & Grade 6 Scores by School Type and Language

Table 1: Average Grade 3 & Grade 6 Scores by School Type and Language (Sorted)

School Type	School Language	Grade 3 Avg	Grade 6 Avg	# of Schools
Public	French	76.66	81.09	76
Catholic	French	71.20	76.67	157
Protestant Separate	English	66.00	78.33	1
Catholic	English	68.49	72.98	980
Public	English	63.24	69.97	1775

This table shows the average academic performance of Grade 3 and Grade 6 across different school types and languages. Based on our data we can see that French-language public schools typically have the highest academic performance. We also see that Catholic French schools perform well, while English Schools (both English and French) do not perform as well with them always being under French Schools. We also note that while Protestant Separate English Schools perform better than English Catholic and English Public Schools, since their population is significantly low (only 1) this result may not be very meaningful or accurate.

4.2 Highest and Lowest 3 Schools Based on Academic Performance

Table 2: Top and Bottom 3 School Boards Based on Academic Performance

Board Name	Schools	Gr3 Avg	Gr6 Avg	% Low-Income	% Parents No Degree
CS Viamonde	39	80.85	84.26	20.00	5.26
CSDC du Centre-Est de l'Ontario	43	79.69	82.05	14.58	2.72
DSB Niagara	74	79.23	80.27	20.42	4.93
DSB Ontario North East	10	35.33	56.00	21.50	8.50
Moosonee DSAB	1	39.00	49.67	35.00	10.00
Moose Factory Island DSAB	1	5.33	32.33	0.00	20.00

This table shows the socioeconomic situation of students who attend schools in the top 3 boards and the lowest 3 boards. When we look at the top 3, we see they have strong scores around the 80% range, while the % of low-income households is moderate, and % of Parents with no High-Level education seems to be low. When we look at the bottom 3, they have severely low results. We also notice that Moose Factory Island DSAB may be an outlier as Gr3 Avg and % of Low-income households significantly differ from the other bottom 3 boards, this may be due to its low # of schools and since we are looking at the extremes. Ignoring the outlier, we see that % of Low-Income households and % of Parents with no High-Level Education are higher than the top 3 counterparts, and also have much less schools.

4.3 Academic Performance by Enrollment Quartile

Table 3: Academic Performance by Enrollment Quartile

Enrollment_Quartile	Gr3 Avg	Gr6 Avg
1	63.02	69.27
2	65.60	71.03
3	66.11	71.76
4	68.14	74.36
NA	69.11	68.67

This table is grouped by the enrollment quartile intervals of all schools. From this table we see a clear positive relationship between enrollment size and academic performance. As when we start from Quartile 1 to Quartile 4, the average score always increases. This pattern is very surprising and goes against many common assumptions.

4.4 Gifted & Special Education Services by Grade Performance Quartiles

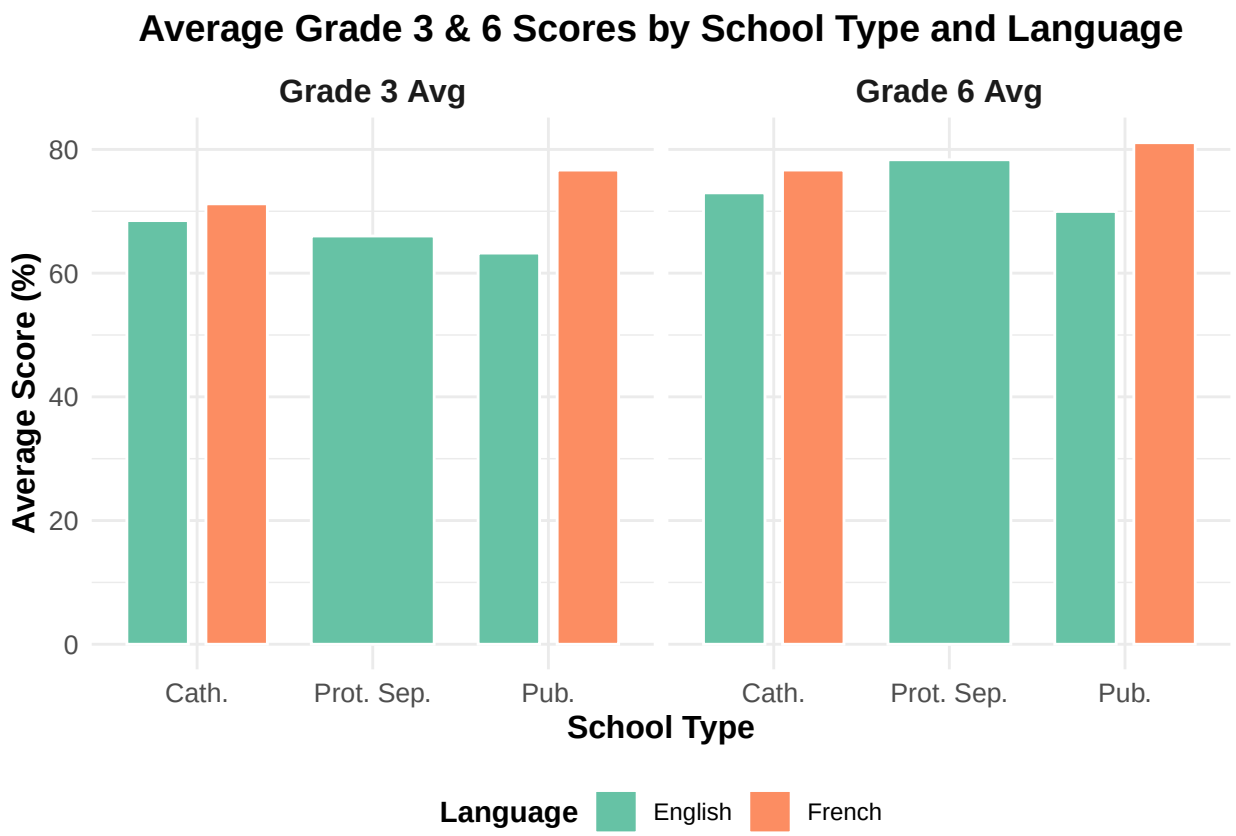
Table 4: Gifted & Special Education Services by Grade Performance Quartiles

Grade_Quartile	Min_Grade	Max_Grade	Mean_Gifted	Mean_SpecialEduServices
1	12.67	60.17	0.13	17.03
2	60.17	70.17	0.18	14.42
3	70.17	78.50	0.88	12.55
4	78.50	98.67	1.14	11.48

This table breaks down schools into quartile groups based on their Grade 3 and 6 academic average combined and shows the average % of gifted students, and students requiring Special Education Services in each interval. We consistently see that schools in the higher grade quartile, consistently have a greater amount of gifted students and a lower amount of students in special education services. Specifically, schools in the 4th quartile have nearly 10x the amount of gifted students in the 1st quartile, while also having about 1/3 fewer students receiving special education services. This pattern could suggest that higher amounts of gifted students and lower amounts of students requiring special education result in better academic performance in school.

5. Graphs

5.1 Double Bar Graph of Student Performance by School Type and Language

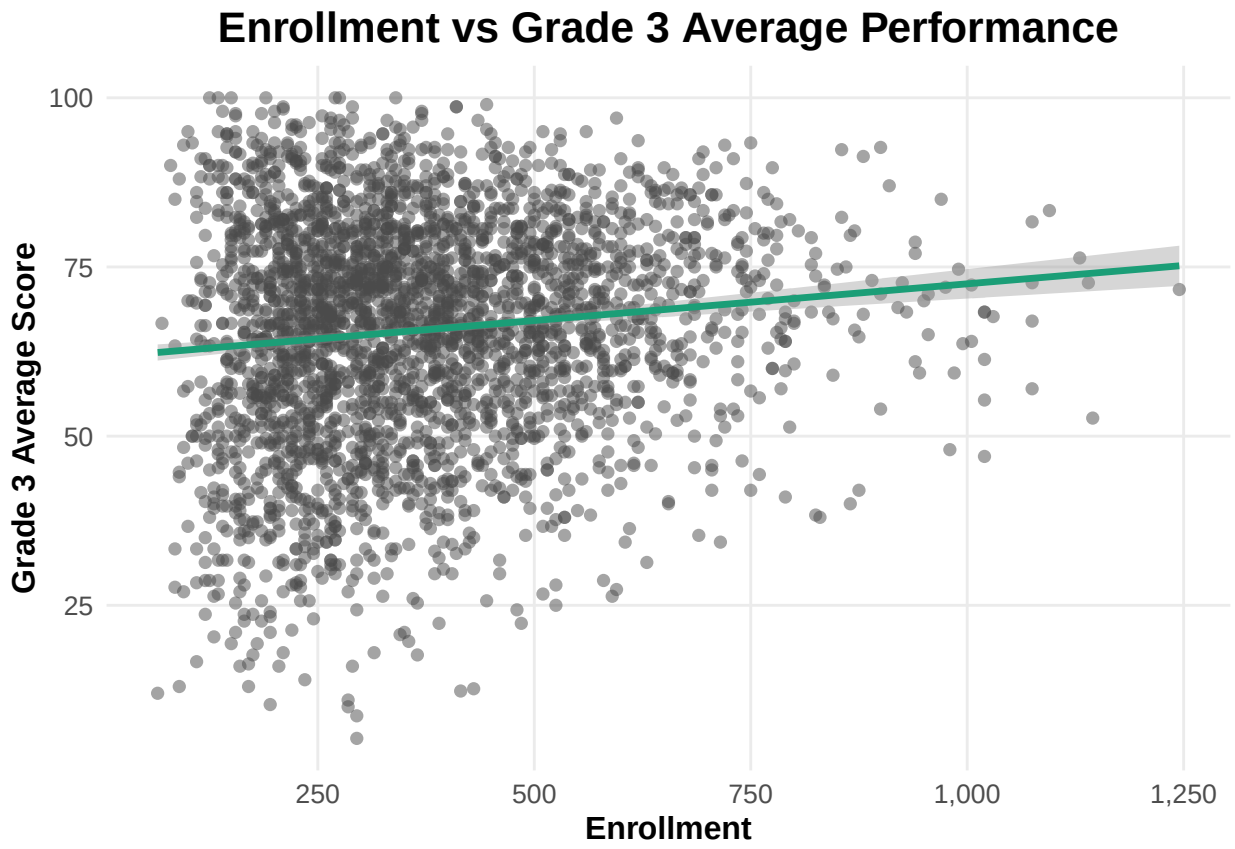


The double bar graph shows the trend of Grade 3 and 6 Performance By School Type and Language for the school information and student demographics data from 2022. The x-axis represents the School Type, which is one of public, catholic, and protestant separate. The y-axis represents the Average Score, in percentage, achieved by students of their respective school type and language. The double bars split the two national languages, English and French, in two different categories. English is in green and French is in orange.

It is important to note that there is only one protestant seperate school in the dataset hence we have an extremely limited understanding of any trends if more such education facilities were to be constructed. The reason it is included is to contrast the more common options of public and catholic school with a more niche and less pushed system. English-language Protestant Separate schools show strong performance in Grade 6, even slightly higher than English Catholic schools. The Protestant Seperate schools also have the greatest growth between grade 3 and 6 for any school type or language in the EQAO. This suggest that, although fewer in number, these schools tend to maintain consistent growth and progress between grades.

A major trend to note is that French-language schools outperform English-language schools Across all school types and both grades. French-language schools consistently score higher than their English counterparts. This is especially evident in public schools, where French Grade 6 scores are significantly higher than English. There is a contrast between English and French speaking public schools. Among English-language schools, public schools have the lowest performance in Grade 3 and Grade 6. However, French-language public schools are the highest performers overall, particularly in Grade 6, possibly indicating stronger academic outcomes in French-language public settings.

5.2 Scatter Plot with Regression Line by Enrollment and Average Grade 3 Scores



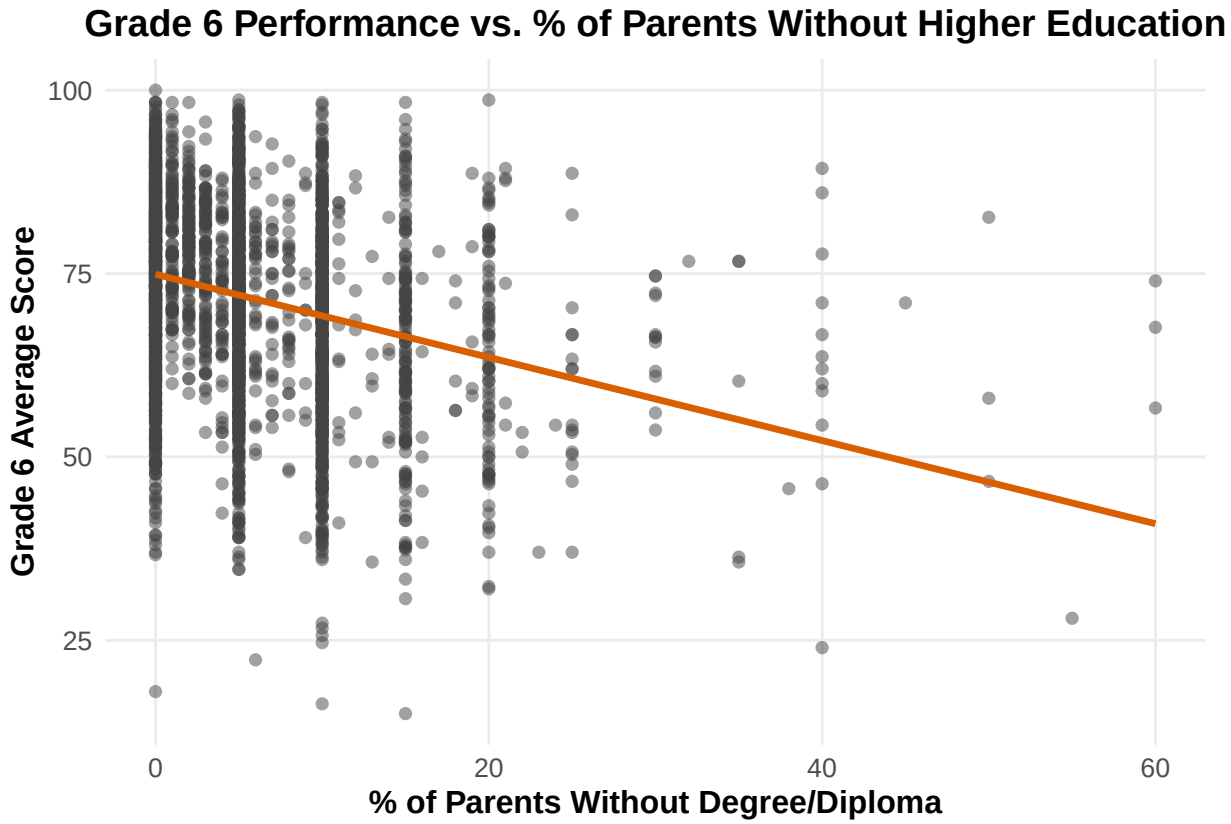
The Scatter Plot of Enrollment Size vs Grade 3 Average Performance displays the average grade 3 performance, calculated by averaging the mean EQAO percentage in math, reading, and writing, for each respective school by the total number of enrolled students in the respective school for the school information and student demographics data from 2022. The x-axis represents the enrollment size of the schools, and the y-axis represents the Average Grade 3 Score, in percentage. There is also a linear regression line with a confidence band that makes the the line of best fit for the trend in the data.

For this graph, the regression line shows a small positive slope, indicating a weak positive correlation between enrollment size and Grade 3 average scores. As enrollment increases, Grade 3 scores tend to slightly improve, though the effect is very marginal. Schools of all sizes display a wide range of academic performance. Some small schools score very high or very low. Large schools also show substantial spread. This suggests that enrollment alone is not a strong predictor of academic outcomes. Most schools are clustered between 100 to 500 students in enrollment.Within this range, the average Grade 3 score typically hovers between 60 and 80.The data suggests that school size is not a strong driver of educational outcomes, reinforcing the need to look beyond structural metrics and explore demographic and socioeconomic variables to better understand student performance.

The confidence band surrounding the regression line represents the range of likely values for the true average relationship between enrollment and performance. From 0 to 750 students, the band is narrowest in the mid-range (200–500) where most schools are concentrated, and wider at the extremes (under 100 and near 750) where data is sparser. This shape reflects the model's greater confidence in predictions where there is more data. The confidence band around the regression line represents the 95% confidence interval, providing a visual cue about the uncertainty in the model's estimated trend. It reminds us that while the regression suggests a slight positive correlation between enrollment and Grade 3 performance, this relationship comes with a degree of statistical variability, especially in areas with fewer schools. The band ensures we interpret

the trend with appropriate caution.

5.3 Scatter Plot With Fitted Regression Line By Parent's Higher Education and Grade 6 Scores

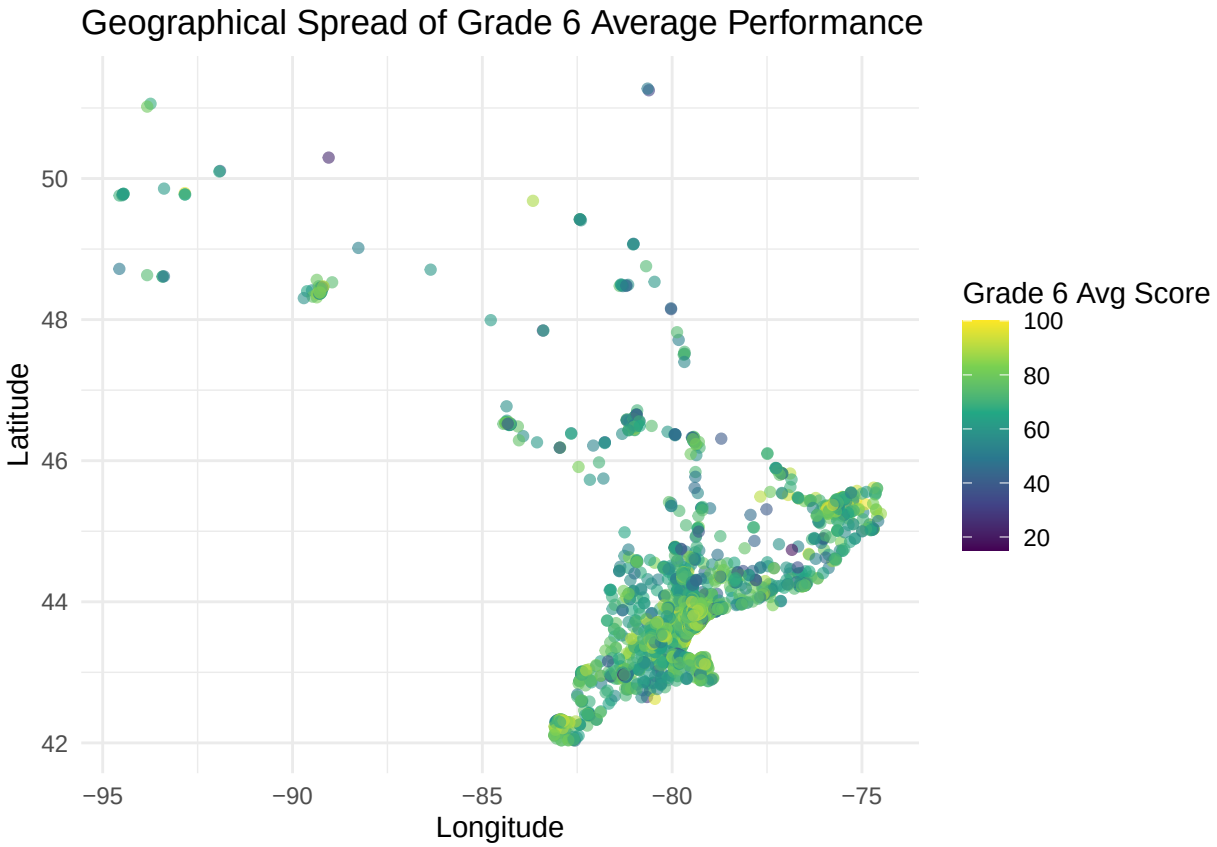


The Scatter Plot of Parents Without Higher Education vs Grade 6 Average Score displays the average grade 6 performance, calculated by averaging the mean EQAO percentage in math, reading, and writing, for each respective school by the total percentage of parents without higher education, such as a degree, diploma, or certificate, for the school information and student demographics data from 2022. The x-axis represents the percentage of the parents in a given school without higher education, and the y-axis represents the Average Grade 6 Score, in percentage. There is also a linear regression line without a confidence band that makes the the line of best fit for the trend in the data.

For this graph, there is a clear negative correlation as the regression line slopes downward, indicating a negative relationship. This indicates as the percentage of parents without higher education increases, Grade 6 scores tend to decrease. This trend is stronger and more visible than in other graphs like enrollment, hence it is useful for a prediction model on this data. Most schools have less than 20% of students whose parents lack a degree. However, even within that lower range, there's still variation in performance, suggesting other influencing factors are also at play. As the percentage of parents without higher education increases, the spread of scores widens. This reflects greater uncertainty and inconsistent outcomes in those schools, possibly due to compounding disadvantages such as, low income or less support at home.

A confidence band was not included in this graph as the there is a clear negative trend between parental education and academic performance present. While confidence bands are useful for showing uncertainty, the core message of this graph that student scores decline as the percentage of parents without higher education increases remains visually strong and easily explicable without it.

5.4 Geographic Scatter Plot By Grade 6 Performance and Lat./Long.



The Geographic Scatter Plot of Grade 6 Average Score displays the average grade 6 performance, calculated by averaging the mean EQAO percentage in math, reading, and writing, for each respective school by geographical location of the school, using longitude and latitude, for the school information and student demographics data from 2022. The x-axis represents the longitude of a given school and the y-axis represents latitude of a given school. There is also a color gradient, which indicates the average grade 6 performance of a given school. It ranges from a dark purple to bright yellow, where the closer it is to the darker shades of purple the lower the performance of the students of the school, while the closer the shade is to a lighter shades of yellow, the higher the performance of the students of the school.

This graph enables us to get a clear spread of the performance of particular school boards and regions and how the education quality differs from region to region. Note, Southern Ontario has the highest school density. The highest concentration of schools is clustered between Latitude 42–45 and Longitude -83 to -75, corresponding to areas like Toronto, Hamilton, Ottawa, and the Golden Horseshoe. These regions also exhibit a diverse range of performance, though many show mid-to-high scores. Schools further north (above Latitude 47) are less frequent and more isolated. Performance in these areas is more variable and tends to skew lower, with some exceptions. There is a noticeable geographic performance divide. Urban/southern regions tend to have higher scores and tighter clustering. While, Rural/northern areas tend to show more variation and lower performance, likely reflecting differences in access to resources, funding, teacher support, and community demographics. These trends emphasize the need to consider location-based educational policy interventions to close performance gaps between regions. Therefore, geographical data should be considered when drawing conclusions on academic performance or making a prediction model.

6. Hypothesis Testing with Confidence Intervals

Analyzing correlation between family income and school performace.

Do schools with more low-income student families perform significantly worse than schools with more high-income student families? This is a research question we aim to test, one which is highly important to understand equity in circumstance of childhood education.

Hypothesis:

- Null Hypothesis (H_0): There is no difference in Grade 6 performance between high-income and low-income groups.
- Alternative Hypothesis (H_1): There is a significant difference in Grade 6 performance between high-income and low-income groups. Schools in low-income groups perform worse (or less likely better), which will be demonstrated in the sample mean of the t-test.

To test this hypothesis, we use the approach of splitting schools into two high-income and low-income groups based on the median percentage of low-income households in each school. We will then compare the students' average Grade 6 scores, and use a two-sample t-test to compare the test scores of the two group of students.

This t-test will give us a p-value, which will indicate whether there is a statistically significant difference averages between the two groups. If this p-value is less than our significance level 0.05 ($p < 0.05$), we reject the null hypothesis. And must conclude there is a significant difference in Grade 6 performance between high-income and low-income groups. Then, we must look the sample means of the high-income and low-income groups to further conclude that schools in low-income groups either perform worse or better than schools in high-income groups.

Finally, we will also calculate the confidence interval, which tells us the range in which the true difference in mean Grade 6 performance between high-income and low-income schools is likely to fall.

Overall, this t-test will allow us to determine if there is a statistical significant correlation between family income and school performance, and will allow us to understand the nature and even percentage amount of this difference.

```
##
##  Welch Two Sample t-test
##
## data:  high_income$`Gr6 Avg` and low_income$`Gr6 Avg`
## t = 15.832, df = 2650, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.657520 8.539787
## sample estimates:
## mean of x mean of y
##  75.07973  67.48107
```

From the t-test results, we see the p-value (2.2e-16) is way smaller than the significance level 0.05. Thus, we must reject the null hypothesis through this statistical evidence, and we accept that there is a significant difference in Grade 6 performance between high-income and low-income school groups. Furthermore, by looking at the sample means of the t-test, we see that the sample mean of x (the high income group's Grade 6 average sample mean) is 75.07973 and the sample mean of y (the low income group's Grade 6 average sample mean) is 67.481077. This shows how the average Grade 6 performance score for each income group differs, showing us that indeed, schools in low-income groups perform worse under the alternate hypothesis. Finally, the confidence interval from the t-test tells us that we are 95% confident that the true difference in

mean Grade 6 performance between high-income and low-income schools falls between 6.66 and 8.54 points. Since zero is not within this range, it confirms that there is a statistically significant difference in academic performance between these groups, providing strong evidence that family income level is associated with academic performance. This means that, on average, students in high-income schools score between 6.66 and 8.54 points higher than those in low-income schools.

7. Bootstrapping

We can use bootstrapping to estimate the average Grade 6 academic performance using sampling, and we can furthermore use a computed confidence interval to understand the usual range of Grade 6 averages.

To do so, we will take repeated samples with replacement of final Grade 6 grades of schools in the data set. The averages of these sampled grades will be used to form a total average for the total number of sampling loops (n_bootstraps) we will run. This will give us the average Grade 6 grade given out throughout all publicly funded schools in the Ontario region.

```
## Mean Grade 6 Mark for Ontario Public School Students: 71.59
```

```
## 95% Confidence Interval for Mean: [ 71.12 , 72.07 ]
```

This result shows that the mean grade 6 mark for Ontario public school students is 71.59333. The 95% confidence interval for this mean is [71.12, 72.07], which means we are 95% confident that the mean grade 6 mark for Ontario public school students is within this range.

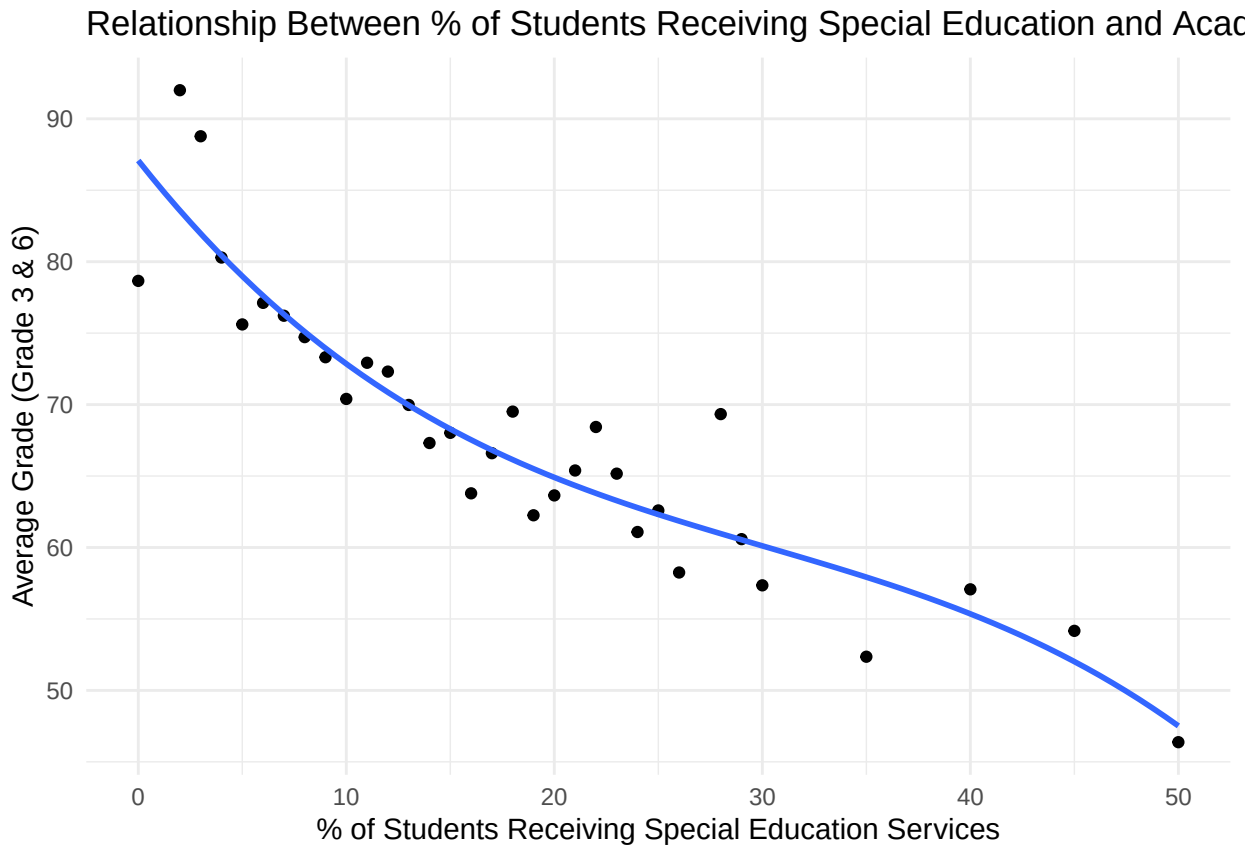
8. Regression Analysis

8.1 Non-Linear Regression Analysis between Special Education Services and Academic Performance

The following model examines how the percentage of students receiving special education services correlates to the overall academic performance of Grade 3 and 6 students. We do this by fitting a cubic regression model to find a potential non-linear relationship between students receiving special education services and academic performance. The result will tell us if there is a correlation and also indicate if there's some sort of threshold where it becomes significantly important or vice versa. .

```
##
## Call:
## lm(formula = Grade_Avg ~ I(`Percentage of Students Receiving Special Education Services`^3) +
##     I(`Percentage of Students Receiving Special Education Services`^2) +
##     `Percentage of Students Receiving Special Education Services`,
##     data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4233 -1.7846 -0.2177  1.4472  8.4171
##
## Coefficients:
##                                                                    Estimate
## (Intercept)                                                      87.0801832
## I(`Percentage of Students Receiving Special Education Services`^3) -0.0005202
```

```
## I(‘Percentage of Students Receiving Special Education Services‘^2)  0.0469923
## ‘Percentage of Students Receiving Special Education Services‘      -1.8405251
##                                                                    Std. Error
## (Intercept)                                                         2.2886693
## I(‘Percentage of Students Receiving Special Education Services‘^3)  0.0002864
## I(‘Percentage of Students Receiving Special Education Services‘^2)  0.0210542
## ‘Percentage of Students Receiving Special Education Services‘       0.4224596
##                                                                    t value
## (Intercept)                                                         38.048
## I(‘Percentage of Students Receiving Special Education Services‘^3)  -1.816
## I(‘Percentage of Students Receiving Special Education Services‘^2)   2.232
## ‘Percentage of Students Receiving Special Education Services‘       -4.357
##                                                                    Pr(>|t|)
## (Intercept)                                                         < 2e-16 ***
## I(‘Percentage of Students Receiving Special Education Services‘^3) 0.079661 .
## I(‘Percentage of Students Receiving Special Education Services‘^2) 0.033507 *
## ‘Percentage of Students Receiving Special Education Services‘      0.000151 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.761 on 29 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.8522
## F-statistic: 62.49 on 3 and 29 DF,  p-value: 9.044e-13
```



The non-linear regression models show us there is a significant relationship between the % of students

receiving special education services and average academic performance. The R-squared of 0.866 shows us that our models can explain 86.6% of the variation in academic performance. The intercept of 87.08 represents the theoretical maximum average performance for if a school had no students receiving special education services. The p-value for the linear (-1.84, p = 0.00015) shows us that initially each 1% increase in special education services is associated with a 1.84 drop in academic performance. The quadratic terms (+0.047, p = 0.033) indicate the negative relationship becomes weaker at % of students receiving special education services. Finally, the cubic term (-0.00052, p = 0.08) is not as significant as our other parameters as its coefficient is small, but it suggests the possibility of subtle long-term effects affecting performance at a significantly high % of students receiving special education services. Together our model suggests that schools experience the most impact from % of students receiving special education services when moving from low to moderate levels, while schools with already high amounts won't have as much of an effect from an increase.

8.2 Linear Regression Analysis Predicting Grade 6 Performance from Socioeconomic and Geographic Factors

This model examines how Grade 6 academic performance is influenced by:

- Percentage of Students Receiving Special Education Services
- Percentage of Parents Without Degrees
- Percentage of School-Aged Children in Low-Income Households
- Percentage of Students Whose First Language Is Not English
- Geographic location (Latitude, Longitude)

The results will identify which of these factors have the strongest statistical relationships with student achievement.

```
##
## Call:
## lm(formula = `Gr6 Avg` ~ `Percentage of Students Receiving Special Education Services` +
##     `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate` +
##     `Percentage of School-Aged Children Who Live in Low-Income Households` +
##     `Percentage of Students Whose First Language Is Not English` +
##     Latitude + Longitude, data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.638  -7.134   0.714   7.782  34.251
##
## Coefficients:
##                                                                                 Estimate
## (Intercept)                                                                   136.89851
## `Percentage of Students Receiving Special Education Services`                  -0.36552
## `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`  -0.24616
## `Percentage of School-Aged Children Who Live in Low-Income Households`         -0.39615
## `Percentage of Students Whose First Language Is Not English`                    0.15164
## Latitude                                                                       -1.15122
## Longitude                                                                       0.05140
##                                                                               Std. Error
## (Intercept)                                                                      9.94063
## `Percentage of Students Receiving Special Education Services`                    0.03578
## `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`    0.03722
```

```
## `Percentage of School-Aged Children Who Live in Low-Income Households`           0.02359
## `Percentage of Students Whose First Language Is Not English`                     0.01059
## Latitude                                                                         0.17783
## Longitude                                                                        0.09165
##                                                                                 t value
## (Intercept)                                                                      13.772
## `Percentage of Students Receiving Special Education Services`                   -10.215
## `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`   -6.613
## `Percentage of School-Aged Children Who Live in Low-Income Households`          -16.792
## `Percentage of Students Whose First Language Is Not English`                     14.324
## Latitude                                                                         -6.474
## Longitude                                                                         0.561
##                                                                                Pr(>|t|)
## (Intercept)                                                                     < 2e-16
## `Percentage of Students Receiving Special Education Services`                   < 2e-16
## `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate` 4.44e-11
## `Percentage of School-Aged Children Who Live in Low-Income Households`          < 2e-16
## `Percentage of Students Whose First Language Is Not English`                    < 2e-16
## Latitude                                                                       1.11e-10
## Longitude                                                                         0.575
##
## (Intercept)                                                                     ***
## `Percentage of Students Receiving Special Education Services`                   ***
## `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate` ***
## `Percentage of School-Aged Children Who Live in Low-Income Households`          ***
## `Percentage of Students Whose First Language Is Not English`                    ***
## Latitude                                                                        ***
## Longitude
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.41 on 2976 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.2765, Adjusted R-squared:  0.275
## F-statistic: 189.6 on 6 and 2976 DF,  p-value: < 2.2e-16
```

This model has R-squared of 0.276 which means its explain to explain 27.6% of the Grade 6 scores with these factors. The overall model is highly significant ($p < 2.2e\text{-}16$), meaning our factors are not enough to give us a accurate prediction.

- Special Education Services (%)
  - Effect: -0.37 points per 1% increase ($p < 0.001$)
  - Interpretation: Lower parental education correlates with reduced performance
- Parents Without Degrees (%)
  - Effect: -0.25 points per 1% increase ($p < 0.001$)
  - Interpretation: Lower parental education correlates with reduced performance
- Low-Income Households (%)
  - Effect: -0.40 points per 1% increase ($p < 0.001$)
  - Interpretation: Strongest negative predictor, more low-income households correlates with reduced performance
- First Language Not English (%)

  - Effect: +0.15 points per 1% increase ($p < 0.001$)
  - Interpretation: Higher % of students whose first language is not English correlates with increased performance

- Latitude
  - Effect: -1.15 points per degree north ($p < 0.001$)
  - Interpretation: School being located further North correlates with reduced performance

- Longitude:
  - Effect: No significant effect ($p = 0.575$)
  - East-west location has very minimal effect on academic performance

9. Cross Validation

Analyzing Grade 6 Academic Performance Based On Key Demographic and Geographic Variables

In this analysis, we can use cross validation to observe the relationship academic performance in Grade 6 and demographic and geographic variables for the school information and student demographics data from 2022. From our findings above, we concluded the factors that were most relevant and important to consider in terms of academic performance were students receiving special education services, students whose parents had no higher education, students from low-income households, students whose first language is not English, and geographical location.

By definition, cross-validation is a model evaluation technique used to test how well a statistical model generalizes to new, unseen data. The idea behind our approach is to apply a 10-fold cross-validation technique to evaluate the performance of a multiple linear regression model that predicts academic performance in Grade 6 based on key demographic and geographic variables. In 10-fold cross-validation, the dataset is split into 10 equal parts, called folds. Then, the model is trained on 9 parts and tested on the remaining 1. This process is repeated 10 times, each time using a different fold as the test set. The performance scores are then averaged to get a final estimate. This helps protect against over fitting and gives us a more reliable estimate of how the model will perform in real-world situations.

The model itself can be evaluated on its accuracy through different measurements, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and lastly R-squared (R^2). RMSE measures the average magnitude of prediction error in the same units as the response variable, Grade 6 score. In this case, lower values are better as they indicate a smaller margin of error in the model prediction. MAE is similar to RMSE but is less sensitive to large outliers hence it provides a more refined and filtered out evaluation. Lastly, R^2 indicates the proportion of variance in Grade 6 Average explained by the model. By definition, values closer to 1 mean a better fit.

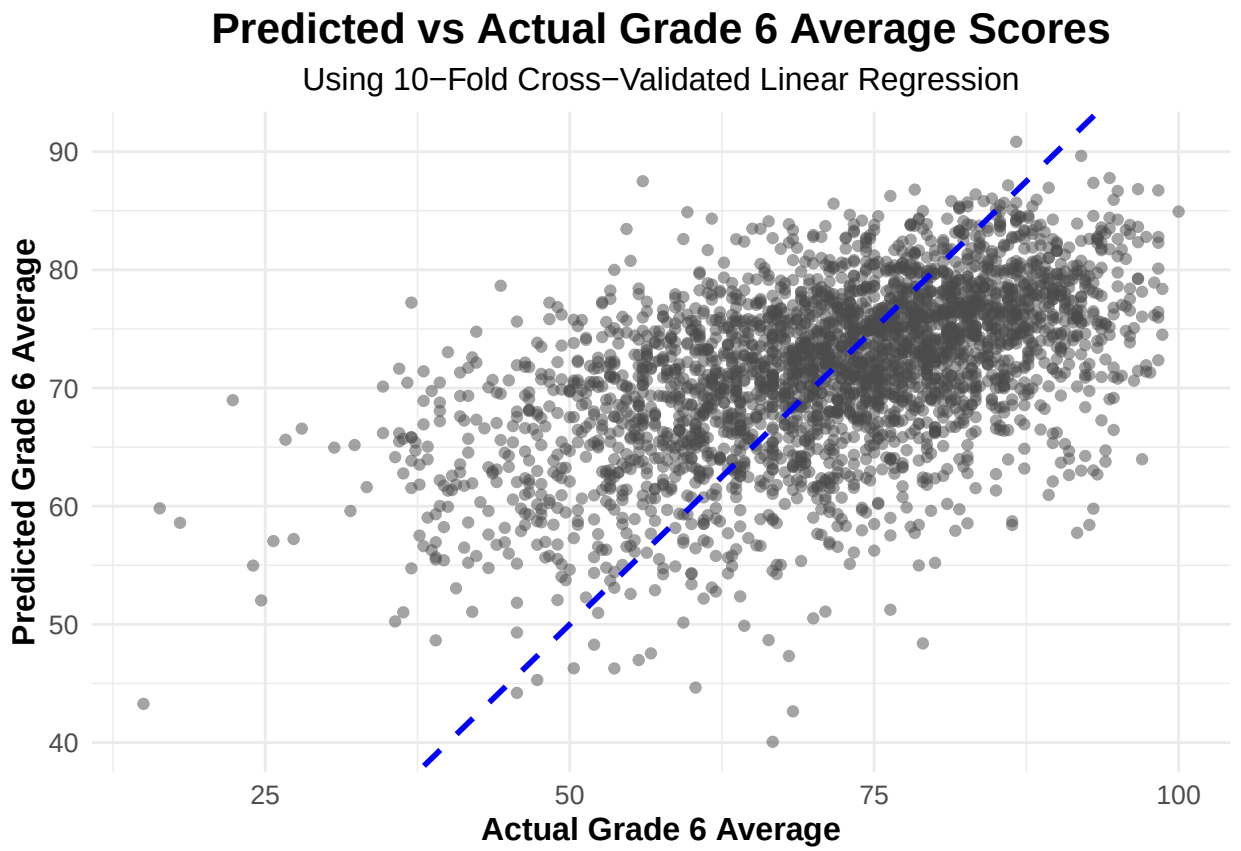
```
## Linear Regression
##
## 2983 samples
##    6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2686, 2686, 2685, 2685, 2685, 2684, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   11.41967  0.2760801  9.011704
```

```
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Conclusion

The average RMSE is around 11.42. The average MAE is around 9.02. The average R^2 is around 0.275. These results suggest that the model's predictions are, on average, ~11.4 points off from the actual Grade 6 average. The refined MAE results suggest a more accurate prediction that is ~9 points off. By the R^2 results, about 27.5% of the variability in academic performance can be explained by the selected variables. While the model captures some meaningful relationships, a large portion of the variation in scores is likely influenced by other unmeasured factors such as, teaching quality, school funding, community resources. This is logically sound, as something as complex and entangled as academic performance of students is much harder to predict as the differendes are too great from person to person. This goes to support that the dataset we used gives a limited understanding of the full effects on the academic performance of elementary schoolers.

Plotting The Result



Using a scatter plot, we plotted the actual vs predicted Grade 6 scores. Ideally, all points would fall directly on the linear regression diagonal dashed line, suggesting a perfect prediction. In our graph, we observe points clustered around the line, which is good. A visible spread confirms moderate predictive power. This supports the earlier conclusion that the model identifies trends, but it doesn't fully capture all the complexity influencing student outcomes which is naturally much more complex.

10. Final Summary

This study examined many factors that influence academic performance for Grade 3 and Grade 6 students in Ontario, leading to the following findings. French language schools consistently had better academic results than their English-language counterparts, with French Catholic schools producing the best academic results. Next we found Socioeconomic factors, specifically a higher percentage of low-income households and parents without higher education, were strongly correlated with lower academic performance. For Geographic locations, we found that schools located farther in the North performed much worse than schools located in the South while schools location West/East had very minimal impacts on academic performance. While enrollment size did not seem to have a significant impact on academic performance we found that a larger enrollment size did correlate to a slight increase in academic performance which was very surprising. As for the overall factors which we found had a significant impact on academic performance we found that a higher percentage of low-income households, more students receiving special education services, and schools located further north were all strong negative predictors of performance. Interestingly, a higher percentage of students whose first language is not English was associated with slightly better academic results. Overall, through our analysis and use of R tools, we were able to make multiple findings on how Socioeconomic status, geographic location and other factors impacted academic performance for elementary students in Ontario.

11. Appendix

```r
knitr::opts_chunk$set(
  echo = TRUE,
  tidy = TRUE,
  tidy.opts = list(width.cutoff = 80)
)
# DATA SETUP

# Importing Libraries
library(tidyverse)
library(readxl)
library(knitr)

# Import Dataset
data <- read_excel("data.xlsx")

# Drop unecessary Columns
data <- data %>%
  select(-contains("Change"),
   -c("Province", "School Number", "P.O. Box", "Building Suite",
      "Street", "Postal Code", "Phone Number", "Fax Number",
      "School Website", "Board Website", "School Special Condition Code",
      "Extract Date", "Board Number", "Board Type", "Municipality",
      "Percentage of Grade 9 Students Achieving the Provincial Standard in Mathematics",
      "Percentage of Students That Passed the Grade 10 OSSLT on Their First Attempt"
      )
   )

# Convert Percent Columns to Numeric (remove "%" sign)
data <- data %>%
  mutate(across(
    c(
```

```
      `Percentage of Grade 3 Students Achieving the Provincial Standard in Mathematics`,
      `Percentage of Grade 3 Students Achieving the Provincial Standard in Reading`,
      `Percentage of Grade 3 Students Achieving the Provincial Standard in Writing`,
      `Percentage of Grade 6 Students Achieving the Provincial Standard in Mathematics`,
      `Percentage of Grade 6 Students Achieving the Provincial Standard in Reading`,
      `Percentage of Grade 6 Students Achieving the Provincial Standard in Writing`,
      `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`,
      `Percentage of School-Aged Children Who Live in Low-Income Households`,
      `Percentage of Students Identified as Gifted`,
      `Percentage of Students Receiving Special Education Services`,
      `Enrolment`,
      `Percentage of Students Whose First Language Is Not English`,
      `Percentage of Students Who Are New to Canada from a Non-English Speaking Country`
    ),
    ~ as.numeric(str_remove(.x, "%"))
  ))


# Compute Grade 3 and Grade 6 Average Scores
d1 <- data %>%
  mutate(
    `Gr3 Avg` = (
      `Percentage of Grade 3 Students Achieving the Provincial Standard in Mathematics` +
      `Percentage of Grade 3 Students Achieving the Provincial Standard in Reading` +
      `Percentage of Grade 3 Students Achieving the Provincial Standard in Writing`
    ) / 3,

    `Gr6 Avg` = (
      `Percentage of Grade 6 Students Achieving the Provincial Standard in Mathematics` +
      `Percentage of Grade 6 Students Achieving the Provincial Standard in Reading` +
      `Percentage of Grade 6 Students Achieving the Provincial Standard in Writing`
    ) / 3
  )

# Remove Rows with Missing Grade 3 or Grade 6 Averages
d1 <- d1 %>%
  filter(!is.na(`Gr3 Avg`), !is.na(`Gr6 Avg`))

# 1.
names(data)
# 4.1

# Show Average Grade 3 & Grade 6 Scores by School Type and Language Table
kable(
  d1 %>%
    group_by(`School Type`, `School Language`) %>%
    summarise(
      `Grade 3 Avg` = mean(`Gr3 Avg`, na.rm = TRUE),
      `Grade 6 Avg` = mean(`Gr6 Avg`, na.rm = TRUE),
      `# of Schools` = n()
    ) %>%
    arrange(desc(`Grade 3 Avg` + `Grade 6 Avg`)),
  caption = "Average Grade 3 & Grade 6 Scores by School Type and Language (Sorted)",
```

```
  digits = 2
)


# 4.2

# Group data and remove N/A rows
board <- d1 %>%
  filter(
    !is.na(`Percentage of School-Aged Children Who Live in Low-Income Households`),
    !is.na(`Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`)
  ) %>%
  group_by(`Board Name`) %>%
  summarise(
    Schools = n(),
    `Gr3 Avg` = mean(`Gr3 Avg`),
    `Gr6 Avg` = mean(`Gr6 Avg`),
    `% Low-Income` = mean(
      `Percentage of School-Aged Children Who Live in Low-Income Households`
    ),
    `% Parents No Degree` = mean(
      `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`
    )
  )

# Show top and bottom 3 boards
top_bottom_boards <- bind_rows(
  board %>%
    arrange(desc(`Gr3 Avg` + `Gr6 Avg`)) %>%
    head(3),
  board %>%
    arrange(`Gr3 Avg` + `Gr6 Avg`) %>%
    head(3)
) %>%
  arrange(desc(`Gr3 Avg` + `Gr6 Avg`))

kable(
  top_bottom_boards,
  caption = "Top and Bottom 3 School Boards Based on Academic Performance",
  digits = 2
)

# 4.3

# Show Academic Performance by Enrollment Quartile Table
enrollment_perf <- d1 %>%
  mutate(Enrollment_Quartile = ntile(Enrolment, 4)) %>%
  group_by(Enrollment_Quartile) %>%
  summarise(
    `Gr3 Avg` = mean(`Gr3 Avg`, na.rm = TRUE),
    `Gr6 Avg` = mean(`Gr6 Avg`, na.rm = TRUE)
  )
```

```r
kable(
  enrollment_perf,
  caption = "Academic Performance by Enrollment Quartile",
  digits = 2
)


# 4.4

# Show Gifted & Special Education Services by Grade Performance Quartiles Table
quartile_table <- d1 %>%
  mutate(Grade_Quartile = ntile(`Gr3 Avg` + `Gr6 Avg`, 4)) %>%
  group_by(Grade_Quartile) %>%
  summarise(
    Min_Grade = min((`Gr3 Avg` + `Gr6 Avg`) / 2, na.rm = TRUE),
    Max_Grade = max((`Gr3 Avg` + `Gr6 Avg`) / 2, na.rm = TRUE),
    Mean_Gifted = mean(
      `Percentage of Students Identified as Gifted`,
      na.rm = TRUE
    ),
    Mean_SpecialEduServices = mean(
      `Percentage of Students Receiving Special Education Services`,
      na.rm = TRUE
    )
  )

kable(
  quartile_table,
  caption = "Gifted & Special Education Services by Grade Performance Quartiles",
  digits = 2
)

#graph1

# Load the required ggplot2 package for data visualization
library(ggplot2)

# Prepare data for the bar plot
plot_data <- d1 %>%
  # Rename long school type names to shorter, cleaner labels
  mutate(`School Type` = recode(`School Type`,
                                "Catholic" = "Cath.",
                                "Protestant Separate" = "Prot. Sep.",
                                "Public" = "Pub.")) %>%
  # Group by school type and language for averaging
  group_by(`School Type`, `School Language`) %>%
  # Calculate average scores for Grade 3 and Grade 6
  summarise(`Grade 3 Avg` = mean(`Gr3 Avg`, na.rm = TRUE),
            `Grade 6 Avg` = mean(`Gr6 Avg`, na.rm = TRUE)) %>%
  # Reshape data to long format for faceted plotting
  pivot_longer(cols = c(`Grade 3 Avg`, `Grade 6 Avg`),
               names_to = "Grade", values_to = "Avg Score")
```

```r
# Create a faceted double bar chart
ggplot(plot_data, aes(x = `School Type`, y = `Avg Score`, fill = `School Language`)) +
  geom_col(position = position_dodge(width = 0.9), width = 0.7, color = "white") +
  facet_wrap(~Grade) +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal(base_family = "Helvetica") +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    strip.text = element_text(size = 12, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "bottom"
  ) +
  labs(
    title = "Average Grade 3 & 6 Scores by School Type and Language",
    x = "School Type",
    y = "Average Score (%)",
    fill = "Language"
  )

#graph2

# Load the required ggplot2 and scales package for data visualization
library(ggplot2)
library(scales)

# Create a scatter plot of enrollment vs Grade 3 average scores
ggplot(d1, aes(x = Enrolment, y = `Gr3 Avg`)) +
  geom_point(alpha = 0.5, color = "#4B4B4B", shape = 16, size = 2) +
  geom_smooth(method = "lm", color = "#1B9E77", linewidth = 1.2) +
  scale_x_continuous(labels = comma) +
  theme_minimal(base_family = "Helvetica") +
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    panel.grid.minor = element_blank()
  ) +
  labs(
    title = "Enrollment vs Grade 3 Average Performance",
    x = "Enrollment",
    y = "Grade 3 Average Score"
  )

#graph3

# Load ggplot2 for data visualization
library(ggplot2)

# Creating plot to show the relationship between parental education and Gr 6 scores
ggplot(d1, aes(
  x = `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`,
```

```
  y = `Gr6 Avg`
)) +
  geom_point(alpha = 0.5, color = "#444444", shape = 16, size = 2) +
  geom_smooth(method = "lm", se = FALSE, color = "#D95F02", linewidth = 1.2) +
  theme_minimal(base_family = "Helvetica") +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    plot.margin = margin(10, 10, 10, 10),
    panel.grid.minor = element_blank()
  ) +
  labs(
    title = "Grade 6 Performance vs. % of Parents Without Higher Education",
    x = "% of Parents Without Degree/Diploma",
    y = "Grade 6 Average Score"
  )

#graph4

# Load ggplot2 for data visualization
library(ggplot2)
# Create a geographic scatter plot showing school locations and Grade 6 performance
ggplot(d1, aes(x = Longitude, y = Latitude, color = `Gr6 Avg`)) +
  # Plot each school as a point with some transparency
  geom_point(alpha = 0.6) +
  scale_color_viridis_c() +
  theme_minimal() +
  labs(title = "Geographical Spread of Grade 6 Average Performance",
       x = "Longitude", y = "Latitude", color = "Grade 6 Avg Score")

# 6. Hypothesis Testing with Confidence Intervals

# Calculate the median low-income percentage
median_income_pct <- median(
  d1$`Percentage of School-Aged Children Who Live in Low-Income Households`,
  na.rm = TRUE
)

# Create income groups
d1_income_grouped <- d1 %>%
  mutate(
    IncomeGroup = ifelse(
      `Percentage of School-Aged Children Who Live in Low-Income Households` >
        median_income_pct,
      "Low-Income",
      "High-Income"
    )
  )

# Split dataset by income group
high_income <- subset(d1_income_grouped, IncomeGroup == "High-Income")
low_income <- subset(d1_income_grouped, IncomeGroup == "Low-Income")
```

```r
# Perform two-sample t-test
t_test_result <- t.test(
  high_income$`Gr6 Avg`,
  low_income$`Gr6 Avg`,
  var.equal = FALSE
)

# View test results
t_test_result


# 7. Bootstrapping

set.seed(123)  # For reproducibility

# Set number of bootstrap samples
n_bootstraps <- 1000

# Generate bootstrap sample means
boot_means <- replicate(n_bootstraps, {
  sample_data <- sample(
    d1$`Gr6 Avg`,
    replace = TRUE
  )
  mean(sample_data, na.rm = TRUE)
})

# Compute 95% confidence interval
boot_ci <- quantile(boot_means, c(0.025, 0.975))

# Display results
cat(
  "Mean Grade 6 Mark for Ontario Public School Students:",
  round(mean(boot_means), 2), "\n"
)
cat(
  "95% Confidence Interval for Mean:",
  "[", round(boot_ci[1], 2), ",", round(boot_ci[2], 2), "]\n"
)


# 8.1

# Grouping by % of students receiving special education and averaging Gr3 & Gr6 performance
d2 <- d1 %>%
  group_by(`Percentage of Students Receiving Special Education Services`) %>%
  summarise(Grade_Avg = (mean(`Gr3 Avg`) + mean(`Gr6 Avg`)) / 2)

# Fitting a cubic regression model
model <- lm(
  Grade_Avg ~
    I(`Percentage of Students Receiving Special Education Services`^3) +
    I(`Percentage of Students Receiving Special Education Services`^2) +
```

```
    `Percentage of Students Receiving Special Education Services`,
  data = d2
)

# Show Model Summary
summary(model)

# Plotting the relationship
ggplot(d2, aes(
  x = `Percentage of Students Receiving Special Education Services`,
  y = Grade_Avg
)) +
  geom_point() +
  stat_smooth(
    method = "lm",
    formula = y ~ poly(x, 3),
    se = FALSE
  ) +
  theme_minimal() +
  labs(
    title =
"Relationship Between % of Students Receiving Special Education and Academic Performance",
    x = "% of Students Receiving Special Education Services",
    y = "Average Grade (Grade 3 & 6)"
  )

# 8.2

# Fitting a linear regression model to predict Grade 6 performance
model <- lm(
  `Gr6 Avg` ~
    `Percentage of Students Receiving Special Education Services` +
    `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate` +
    `Percentage of School-Aged Children Who Live in Low-Income Households` +
    `Percentage of Students Whose First Language Is Not English` +
    Latitude +
    Longitude,
  data = d1
)

# Show model
summary(model)

# Load caret for machine learning and cross-validation tools
library(caret)

# Prepare the data for modeling
cv_data <- d1 %>%
  # Select the outcome variable and relevant predictors
  select(`Gr6 Avg`,
         `Percentage of Students Receiving Special Education Services`,
         `Percentage of Students Whose Parents Have No Degree, Diploma or Certificate`,
         `Percentage of School-Aged Children Who Live in Low-Income Households`,
```

```r
         `Percentage of Students Whose First Language Is Not English`,
         Latitude, Longitude) %>% drop_na()

# Set up the 10-fold cross-validation procedure
cv_control <- trainControl(method = "cv", number = 10)

# Train the linear regression model using cross-validation
cv_model <- train(
  `Gr6 Avg` ~ .,
  data = cv_data,
  method = "lm",
  trControl = cv_control
)

# Output cross-validation results (RMSE, R², MAE)
cv_model
# Create a data frame that compares actual vs predicted Grade 6 scores
cv_predictions <- data.frame(
  Actual = cv_data$`Gr6 Avg`,
  Predicted = predict(cv_model, newdata = cv_data)
)

# Create a scatter plot to compare actual vs predicted values
ggplot(cv_predictions, aes(x = Actual, y = Predicted)) +
  geom_point(alpha = 0.5, color = "#4B4B4B") +  # Plot each school as a point
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "blue", linewidth = 1) +  # Refere
  theme_minimal(base_family = "Helvetica") +
  labs(
    title = "Predicted vs Actual Grade 6 Average Scores",
    subtitle = "Using 10-Fold Cross-Validated Linear Regression",
    x = "Actual Grade 6 Average",
    y = "Predicted Grade 6 Average"
  ) +
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5),
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10)
  )
```