Final Project

Hongshuo Zhou

Contents

1 Introduction **2**
1.1 Research Question 2

2 Methods **2**
2.1 Data Collection 2
2.2 Data Wrangling 2
2.3 Data Summary 3
2.4 Data Visualization 4
2.5 Tools Used 5

3 Results **6**
3.1 Summary Visuals 6
3.2 Modelling 7

4 Conclusion **11**
4.1 Limitations and Future Directions 11

1 Introduction

Education is one of the fundamental building blocks of any society and plays a vital role in shaping the future of individuals, communities, and nations. Among all years in school, k-12 education has the most significant influence on students because their personality and vision of the world are primarily shaped in elementary and secondary schools. Factors such as the school's suspension rate, drop rate, number of students, full-time educators, and geographic locations can partially reflect a school's academic performance. For example, suspension rates in primary and secondary schools are of concern as they are associated with adverse outcomes such as reduced academic performance, increased dropout rates, and a higher likelihood of involvement in delinquent behavior. On the other hand, other factors like the teacher-student ratio can suggest how many school resources and attention teachers each student can get. Thus a smaller ratio might imply a higher academic performance of a school board. Hence, from these factors, we can analyze their impacts on a school's academic performance, and understanding the relationship between them is crucial for policymakers, educators, and stakeholders in designing effective strategies to improve educational outcomes.

With this in mind, I want to investigate the following question: What factors of a school will influence their academic performance and their relationships? To be more specific, what are the relationships between each school board's academic performance and their suspension rate, drop rate, number of students, number of full-time educators, and geographic locations?

The data I obtained to answer the questions is from Ontario public data website and Open Canada searching with keyword *school board*. The datasets contain information about all public schools, including elementary and secondary schools, in Ontario. All datasets I can find are from 2015 to 2019, so I will also limit my scope to this range.

1.1 Research Question

What factors of a school will influence their academic performance and their relationships? To be more specific, what are the relationships between each school board's academic performance and their suspension rate, drop rate, number of students, number of full-time educators, and geographic locations?

2 Methods

2.1 Data Collection

We obtain all our datasets from Ontario public data website and Open Canada by searching for the keyword `school board`. To be more specific, we download datasets for information on each school board's academic standing, each school board's full-time educators, each school board's students, each school board's schools, each school board's expulsion rate, and each school board's suspension rate.

We download all the datasets mentioned above and load than using `read_excel`. Since datasets for academic performance, educators, and students have their data stored in a separate file for different years, we simply use `rbind` to combine them into a single table. In addition, because the datasets of expulsion and suspension rates have a format that each roll is a school board and each column is the suspension rate for a specific year, we created two new datasets to make them into the correctly formatted datasets.

2.2 Data Wrangling

Since the original datasets collected are too messy to summarize and display, I will perform data cleaning and wrangling first and then show the cleaned form of the dataset.

2.2.1 Data Cleaning

We drop all the irrelevant variables that do not provide insight into the question of interest, such as variable `phone number` in the dataset of schools information, variable `Four Year Graduation Rate` in the dataset for academic performance, and some categorical variables in datasets for expulsion rate and suspension rate.

We also notice that for enrollment status datasets, despite all containing the same information, the column names are different across datasets, and thus we rename them to the same name for future simplicity. Next, since the academic performance datasets contain information for both elementary and secondary schools identified by the variable `School Level`, I separate the single dataset into two dataframes, one for elementary and another for secondary.

In addition, the passing rates for EQAO and OSSLT are documented as characters with the format `x%`, and the elementary male and female enrollment numbers and secondary male and female enrollment numbers are as characters even though they are numbers. Hence, I convert them to numerical variables, which are more intuitive.

2.2.2 Data Merging

For elementary and secondary datasets each, we merge them with other five tables to get two final datasets that will be used in the analysis. During the merging, we ensure common variables `Board Number`, `Start Year`, `School Name` are matched for each row every time we merge two datasets. After taking a closer look at the two merged datasets, we detected that elementary contains 5 missing values and secondary contains 8 missing values, so we dropped all rows with missing values.

2.2.3 Feature Engineering

We want to create two variables `Elementary Student-Teacher Ratio` and `Secondary Student-Teacher Ratio`, which is the number of students in the school board divided by the number of teachers in the row. We create this using function `mutate` and the two new variables are calculated as previously mentioned. We also created a categorical variable, `Passing rate category`. This column is made based on the numerical passing rate by comparing each of them to the mean passing rate across the whole dataset. If it is at least the mean, then we document it as `high` and `low` otherwise.

Now, we obtain the final datasets that we can work with. The elementary school information dataset has 15,003 rows and 25 columns, and the secondary school information dataset has 3414 rows and 25 variables.

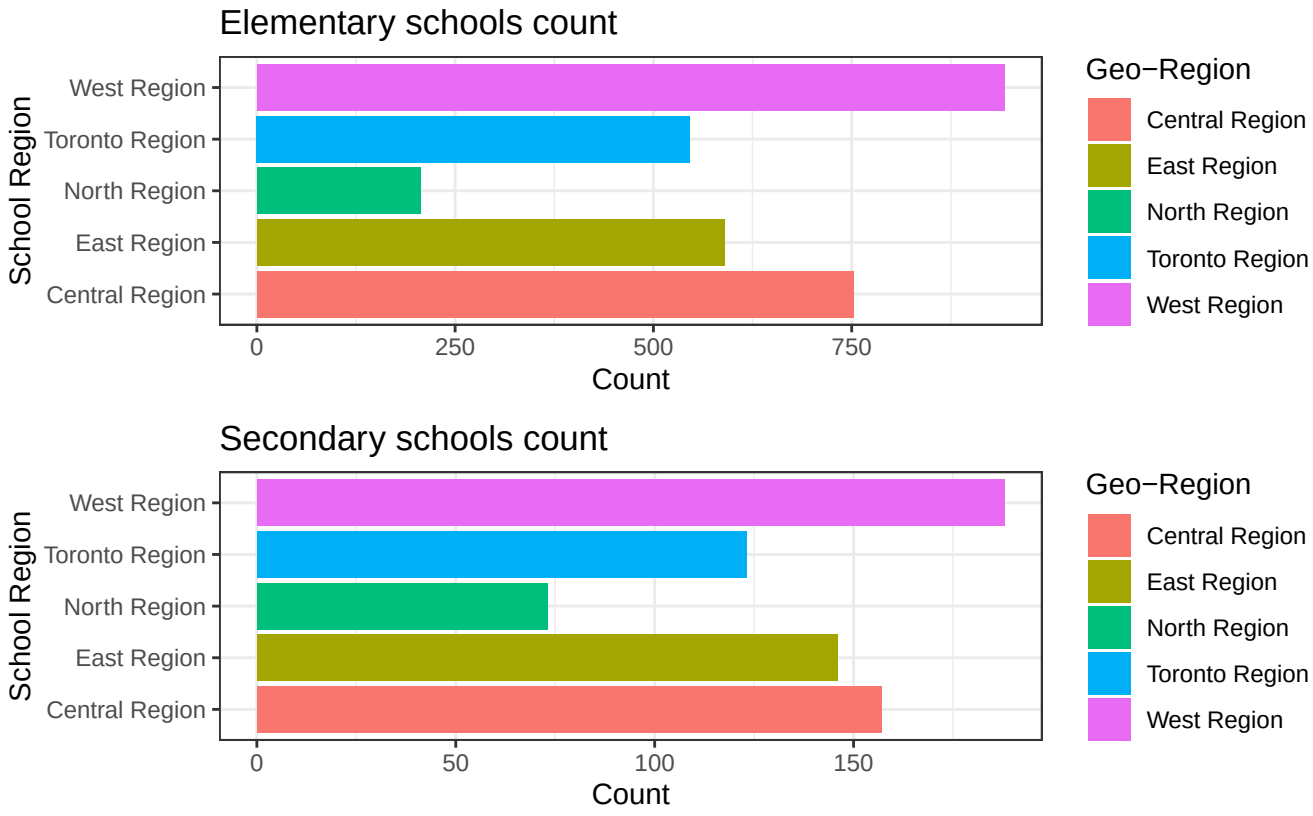
2.3 Data Summary

Table below is a description of the crucial variables that we used in this report. We only include part of the variables since others such as `Latitude` and `Longitude` are just for visualization not analysis.

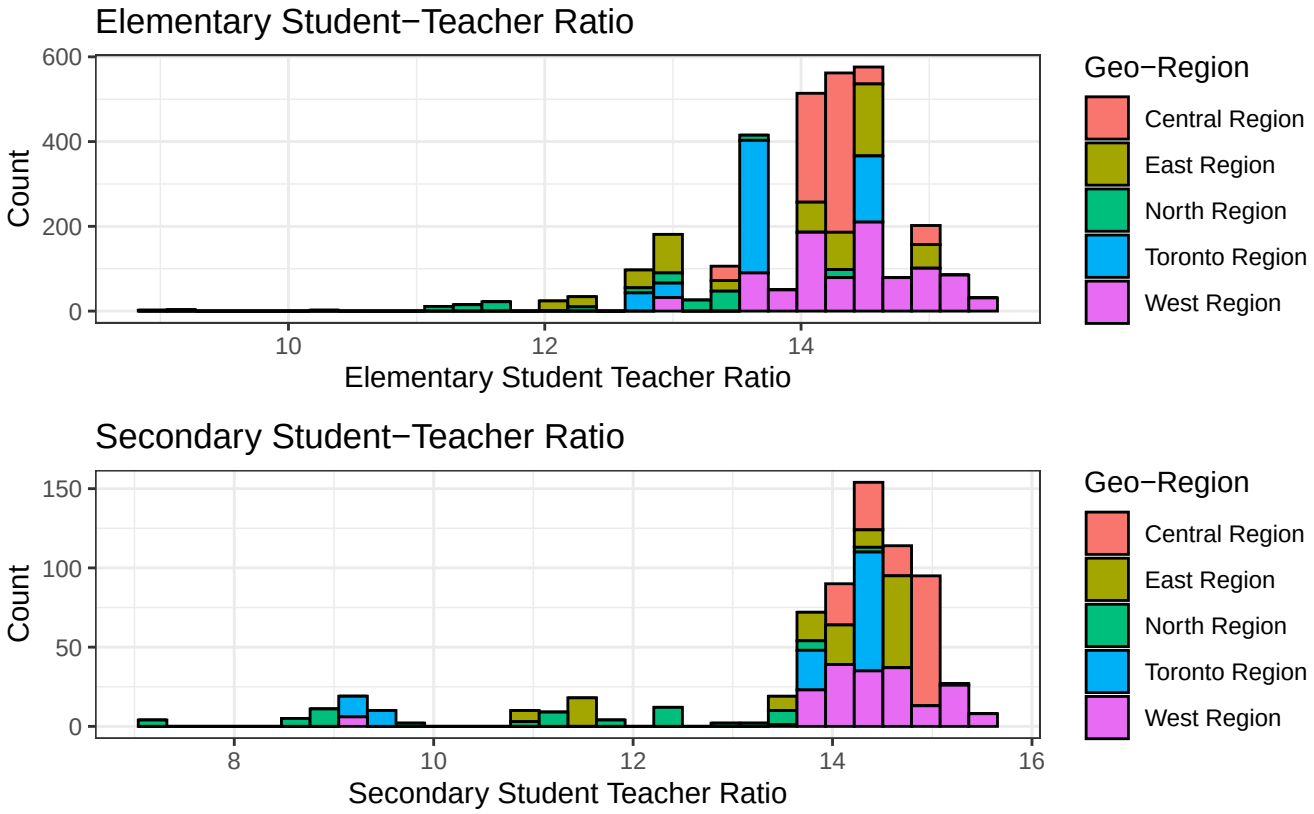
Variable Name	Variable Description
Board Name	The name of the school board.
Board City	The city of the school board in.
Grade 6 EQAO Reading Results	The passing rate of the EQAO reading test.
Grade 10 OSSLT Results	The passing rate of the OSSLT test.
Start Year	The start year of the school year.
Elementary Male Enrolment	The number of elementary school male students enrolled.
Elementary Female Enrolment	The number of elementary school female students enrolled.
Secondary Male Enrolment	The number of secondary school male students enrolled.
Secondary Female Enrolment	The number of secondary school female students enrolled.
Expulsion Rate	The expulsion rate of the school board in the year.
Suspension Rate	The suspension rate of the school board in the year.
Elementary Student Teacher Ratio	The ratio between elementary student and elementary full-time educators.
Secondary Student Teacher Ratio	The ratio between secondary student and elementary full-time educators.

2.4 Data Visualization

To better understand the dataset, we decided to explore some of the variables that could have an impact on our response variable, academic performance. After exploring different variables and plots, we find the number of schools in each region and the distribution of the student-teacher ratio has the most interesting patterns. They are examined by splitting on the school level.



The barplots show the number of schools in each region. As we can see, the proportion of schools in each area is the same across primary and high schools; the west region has the most schools, and the north region has the least. We can see that the data is reasonable because the northern part of Ontario is too cold that the number of habitants is low compared to the other four parts of Ontario. However, we notice many more elementary than secondary schools in all five areas. Hence, we try to find supporting data, which says in Ontario, there are roughly 4000 elementary schools and 900 high schools. This coincides with our data, considering some schools have missing values.



The histograms are for the student-teacher ratio in each region. According to the graph, both two distributions have a very heavy left tail. The center of a high school's Student-Teacher Ratio is approximately 14, and the center of an elementary school's Student-Teach Ratio is 13. From this, we can conclude that, on average, elementary schools have a lower student-teach ratio, and by our definition of the term, we would conclude that elementary school students have better access to the schools' resources. Another possible explanation is that it might be because teenagers in secondary schools do not need that much attention compared to younger students in primary schools. Also, we notice that the student-teacher ratios for the west area are on the upper tail, whereas those for the north region are on the lower side. We find this matches the number of schools in each region, suggesting regions that have more schools also have more students, and more schools don't mean better access to teaching resources.

Lastly, we use `leaflet` to map all the schools, and we see that the number of elementary schools is indeed larger than the number of secondary schools. The maps can be seen on the website.

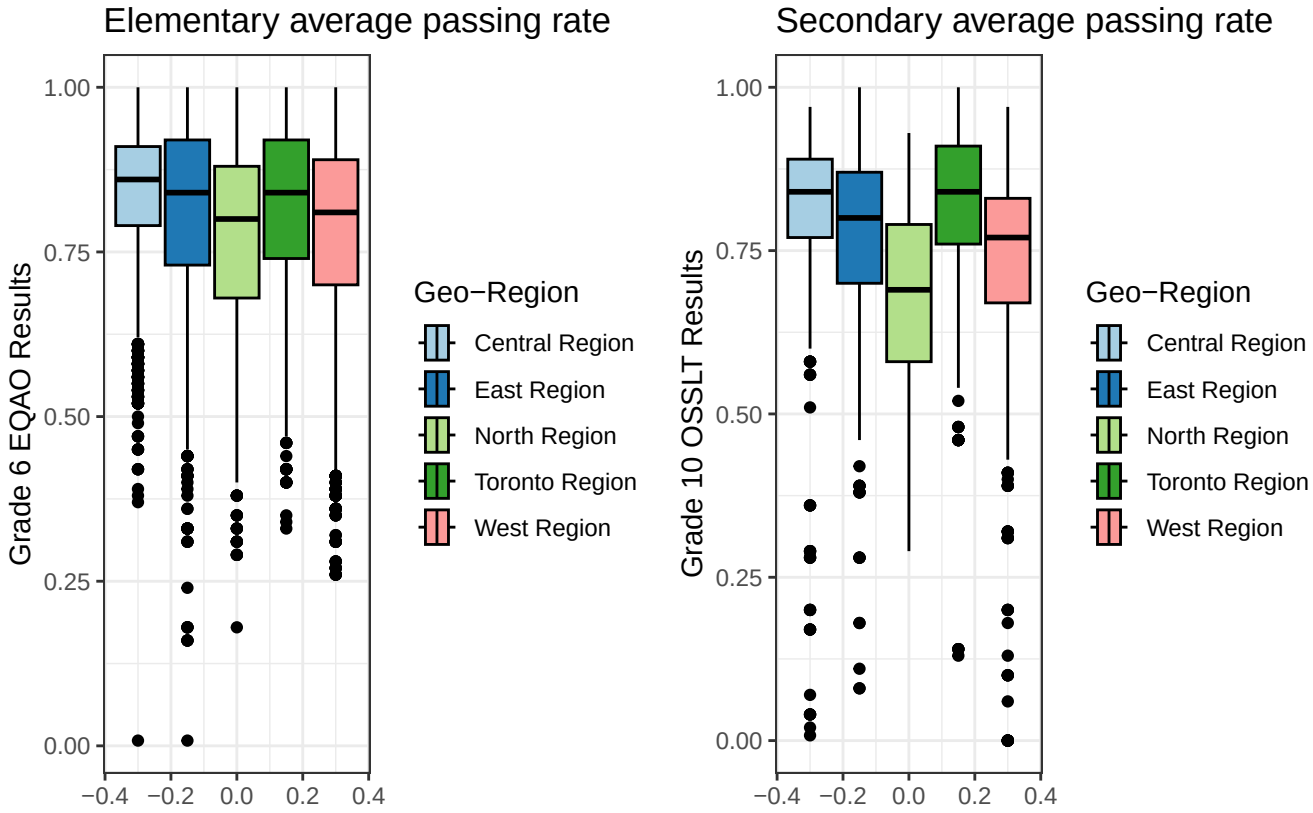
2.5 Tools Used

The tools used in data wrangling are from packages `tidyverse` and `dplyr`. All tables are created with `kable`. Figures and plots are made using `ggplot2`, interactive visuals use package `plotly`, and maps are created with `leaflet`. The models we use are `lm`, and `gam`.

3 Results

3.1 Summary Visuals

3.1.1 Spacial Visuals

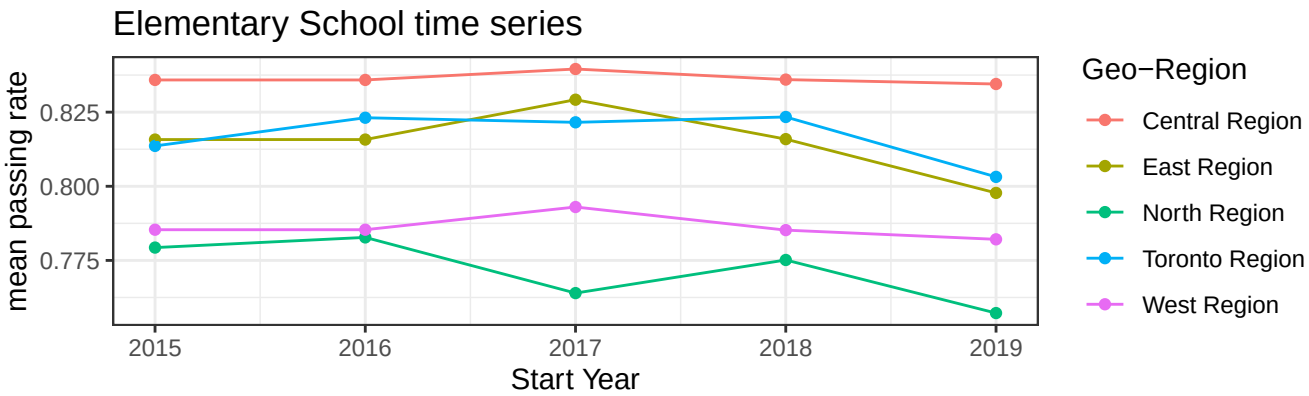


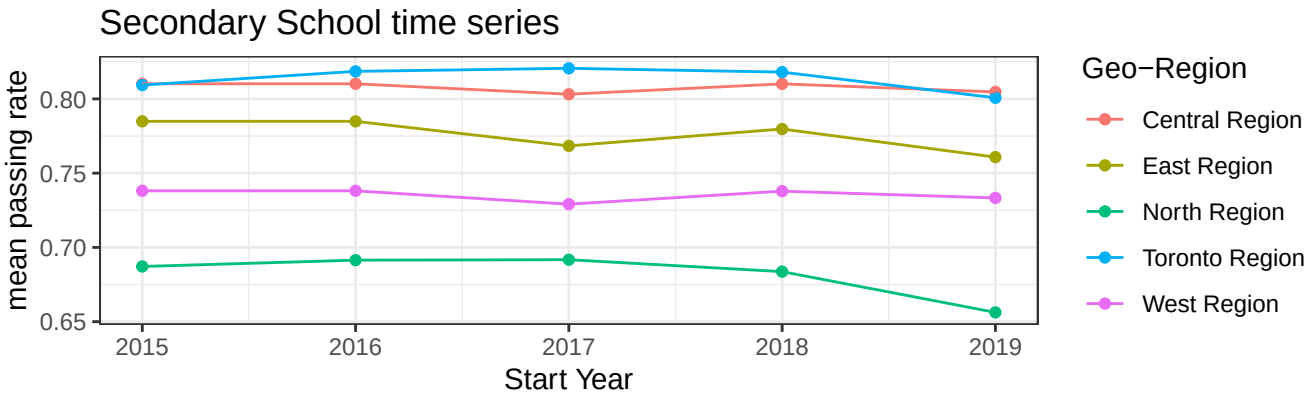
Due to the difference in the number of schools in each region, as shown above, it is natural to investigate how the geographical area of a school board affects exam passing rates. Hence, we plot the boxplot of elementary and secondary school's academic performance in each region. For the elementary schools' boxplot, the medians of five regions are relatively the same, which is around 0.85. However, the medians of secondary schools are far more separated compared to the previous plot. Among all five areas, the Toronto region and the central region have the highest passing rate, whereas the northern region is the lowest. Both school levels have a heavy tail on the lower end, and some schools even have passing rates of less than 20 percent.

One possible explanation for the larger dispersion between the medians for secondary school academic performance is that the materials in high schools are more challenging and harder to understand. During elementary school, kids normally learn the basic concepts, which suggests more kids can master the concepts and receive a better passing rate. However, some courses in high school, such as physics, chemistry, and literature, require more hard work. So, the region, or the quality of resources, plays a more prominent role in the school boards' academic performance, and thus the medians begin to split. Moreover, the odd passing rate at the lower end might be a documentation error, but I don't have any information to back this up. Therefore, I choose to trust the data and keep them.

3.1.2 Temporal Visuals

Since we have the data across five years, the effects of time on academic performance can also be analyzed. So, we decide to make line plots for elementary and secondary school average academic performance.





As we can see, the averages are approximately the same during the five years with some minor fluctuations. Similar to the previous boxplots, the northern region has the lowest mean, and the central and the Toronto region has the highest average passing rate. Furthermore, for each region, the passing rate of the grade 6 EQAO exam is higher than the passing rate of the grade 10 OSSLT exam.

3.2 Modelling

3.2.1 Linear Regression

First, we note that our variable of interest is a numerical value, and based on the previous visualization, the predictors seems to have a linear relationship with the academic performance. Under these conditions, we would have considered using a a linear model to answer the question. The models are shown below:

Table 2: Coefficients for Elementary Schools' Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.6755922	1.4932611	3.800804	0.0001448
`Start Year`	-0.0022663	0.0007402	-3.061775	0.0022042
`Geo-Region`East Region	-0.0034566	0.0033547	-1.030371	0.3028526
`Geo-Region`North Region	-0.0435348	0.0055211	-7.885205	0.0000000
`Geo-Region`Toronto Region	-0.0317395	0.0035891	-8.843291	0.0000000
`Geo-Region`West Region	-0.0141507	0.0030958	-4.570948	0.0000049
`Expulsion Rate`	-23.5104952	5.6562276	-4.156568	0.0000325
`Suspension Rate`	-2.2272115	0.0906026	-24.582203	0.0000000
`Elementary Student Teacher Ratio`	-0.0152511	0.0017924	-8.508770	0.0000000

Table 3: Coefficients for Secondary Schools' Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.7045758	3.3790553	0.8003941	0.4235384
`Start Year`	-0.0008053	0.0016769	-0.4802241	0.6310989
`Geo-Region`East Region	-0.0009087	0.0075665	-0.1200907	0.9044184
`Geo-Region`North Region	-0.0991150	0.0111458	-8.8925659	0.0000000
`Geo-Region`Toronto Region	-0.0234247	0.0079621	-2.9420240	0.0032827
`Geo-Region`West Region	-0.0268891	0.0072328	-3.7176476	0.0002043
`Expulsion Rate`	19.7960659	12.7979997	1.5468094	0.1220021
`Suspension Rate`	-3.1358838	0.1894189	-16.5552879	0.0000000
`Secondary Student Teacher Ratio`	-0.0141432	0.0018322	-7.7192126	0.0000000

3.2.2 Model Interpretation

For the elementary school's linear model, we can see the only coefficient that is not significant is for the east region. It means, on average, the passing rate for the east region is the same as the passing rate for the central region while holding all other variables constant. Since all region dummy variables' coefficients are negative, it solidifies our observation that central region schools have the highest elementary academic performance. We also notice that the coefficients for the `expulsion rate`, `suspension rate`, and `student teacher ratio` are negative. This indicates higher these variables, the poorer the academic performance of a school that meets our expectations. The R^2 for this model is roughly 0.1, suggesting that our model only explains 10 percent of the variation.

The second model is the high school's linear model. This time, only coefficients for `North Region`, `Toronto Region`, `West Region`, `Suspension Rate`, and `student teacher ratio` are significant. Still, the coefficients are all negative, and the interpretations are exactly the same as the elementary school's linear model. The R^2 for this model is 0.16, suggesting that our model only explains 16 percent of the variation. Due to the poor performance of these two models, we decide the try other models to see if we can improve their performance.

3.2.3 Generalized Additive Model

Since our variable of interest is a numerical variable, another model we learned in the lecture is the generalized additive model. Since the passing rates should follow a normal distribution, we use the Gaussian distribution as the family.

Table 4: Coefficients for Elementary Schools' Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0789417	0.0256992	41.9834278	0.0000000
GeoRegionEast Region	-0.0008143	0.0033596	-0.2423622	0.8085029
GeoRegionNorth Region	-0.0366793	0.0055655	-6.5904572	0.0000000
GeoRegionToronto Region	-0.0322222	0.0035811	-8.9979242	0.0000000
GeoRegionWest Region	-0.0117490	0.0030994	-3.7907681	0.0001508
ExpulsionRate	-23.0299766	5.6460268	-4.0789705	0.0000455
SuspensionRate	-2.4351603	0.0933925	-26.0744668	0.0000000
StudentTeacherRatio	-0.0131439	0.0018066	-7.2755550	0.0000000

Table 5: Coefficients for Secondary Schools' Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0643906	0.0280524	37.9429057	0.0000000
GeoRegionEast Region	0.0014888	0.0075983	0.1959355	0.8446724
GeoRegionNorth Region	-0.0931272	0.0113103	-8.2338248	0.0000000
GeoRegionToronto Region	-0.0231121	0.0079531	-2.9060429	0.0036839
GeoRegionWest Region	-0.0246586	0.0072614	-3.3958267	0.0006921
ExpulsionRate	20.3619826	12.7876771	1.5923128	0.1114073
SuspensionRate	-3.2661463	0.1941059	-16.8266234	0.0000000
StudentTeacherRatio	-0.0128725	0.0018793	-6.8498188	0.0000000

3.2.4 Model Interpretation

Looking at the elementary school's gam model, we notice only the term for the east region is insignificant. This coincides with our simple linear model, suggesting that, on average, the passing rate for the east area is the same as the passing rate for the central region while holding all other variables constant. Similarly, all region's dummy variables' coefficients are negative, suggesting the central area has the highest academic performance. As for variables `expulsion rate`, `suspension rate`, and `student teacher ratio`, they are still negative and suggest they are negatively related to the response variable. The R^2 for this model is approximately 0.12, offering that our model only explains 12 percent of the variation. Overall, compared to the linear model, the improvement is relatively small.

Then, let's closely examine the high school's gam model. The expulsion rate is also insignificant in this model besides the east region. All significant coefficients are negative, which is similar to the elementary school's gam model. The R^2 for this model is approximately 0.20, offering that our model only explains 20 percent of the variation. The improvement compared to linear regression is still not big.

3.2.5 Model Comparison

Lastly, we perform anova test to compare the linear model and gam model and decide which one we should pick.

Table 6: Elementary Models ANOVA

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
14994.00	243.1891	NA	NA	NA	NA
14991.53	241.8135	2.471485	1.375542	34.50494	0

Table 7: Secondary Models ANOVA

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3405.000	63.39767	NA	NA	NA	NA
3402.679	63.19528	2.320782	0.202385	4.695485	0.006283

Since the p-value for two primary school models is less than 0.05, we should reject the null hypothesis. Hence, for the elementary schools, we should use the gam model. However, the p-value for the two high school models is greater than 0.05, indicating that the linear model is more suitable and the improvement of the gam model is insignificant.

In short, among the models we fitted, the linear model is more suitable for the analysis of secondary school

academic performance, and it has an R^2 of 0.16. For analysis of elementary schools' academic performance, it's better to pick the gam model, which has an R^2 of 0.12. Overall, the explanatory power of the two models is quite poor.

4 Conclusion

From spacial analysis, we find out that the geographical region does affect the academic performance of a school. This pattern is more obvious in the plot for secondary school exam passing rates, and among all five areas, the central region always achieves the highest academic performance, whereas the northern region is the lowest. In the temporal analysis, we conclude that the mean academic performance of each region remains constant with minor fluctuations. This might be due to the short time span of our dataset, which only includes five years of data. When consulting outside sources, some paper points out that the passing rates are declining for both elementary and secondary schools if we look at a period of 20 years. Then, we fit linear models for the two school levels. The results indicate a linear relationship between the response variable and the predictors. The explanatory power, nevertheless, is not desirable for both of them since all two R^2s are lower than 0.2. To find models that raise the explanatory power, we fit gam models. Even though the explanatory power increased for school levels, the outcomes are still undesirable. Finally, we use ANOVA to compare the two models and find out whether the linear model is more suitable for analyzing secondary school academic performance and whether it's better to pick the gam model for secondary schools. From the models, we are able to say that if only considering the geographic effects, the central region has the highest average academic performance. Also, the `expulsion rate`, `student-teacher ratio`, and `suspension rate` are negatively related to academic performance and meet our prediction before the analysis.

4.1 Limitations and Future Directions

There are several limitations to the current study. One of the biggest problems is insufficient data; especially we only have five years of data. Because of this, we cannot really explore how the student's academic performance differs over the years, which has been shown in various other studies. So, even though the year variable is significant in the model, it doesn't add much insight into the question. Hence, in the future, we can find more data to improve the models. Additionally, our models' performance is lower than the expectation, which explains at most 20 percent of the variation of the data. Even though the results meet the intuition and predictions, we still need to strive for models with better explanatory powers. Therefore, in the future, we can try other more complex models, such as neural networks, to better analyze the patterns and provide more insights into the question.