Final Project

Hongshuo Zhou

Contents

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. Now, we obtain the final datasets that we can work with. The elementary school information dataset has 15,003 rows and 24 columns, and the secondary school information dataset has 3414 rows and 24 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

Figure 1 − A barplot of the number of schools in each region

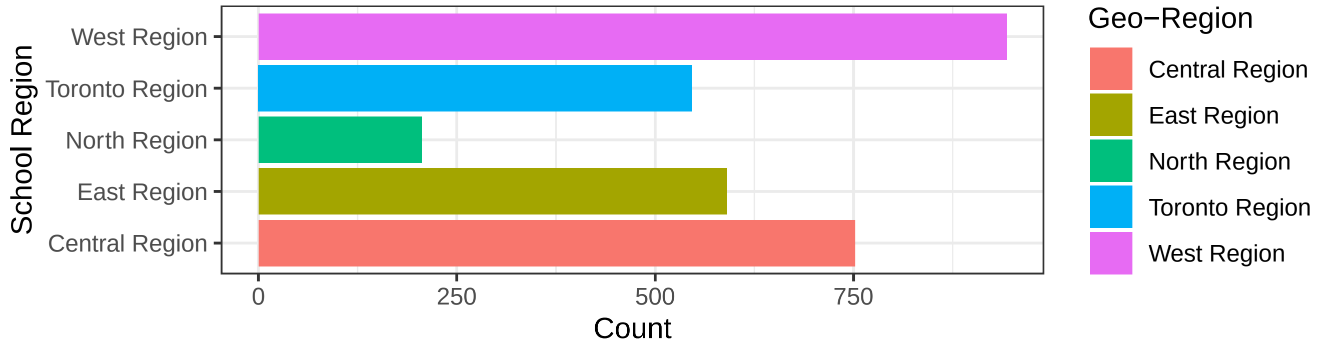
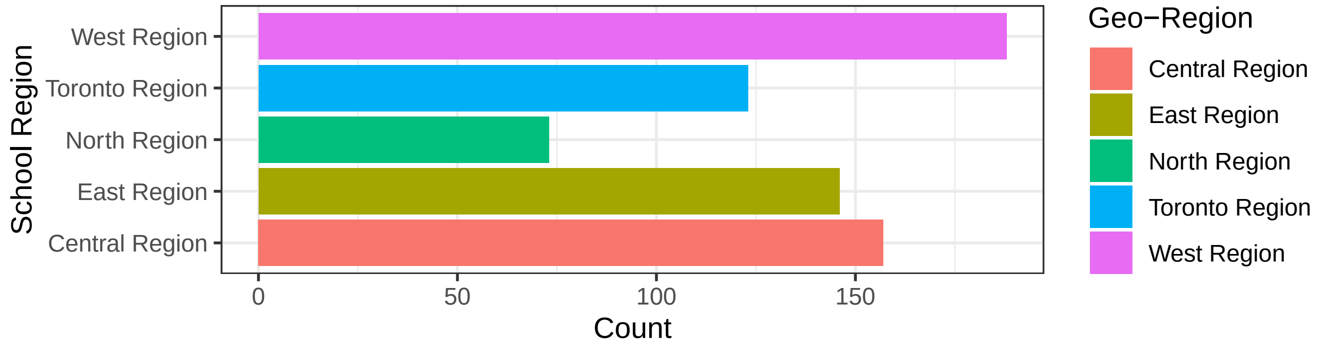


Figure 1 − A barplot of the number of schools in each region



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.

3 Results

3.1 Summary Visuals

3.2 Modelling

3.3 Classification

4 Conclusion