School After Recess: Exploring School Closures (2012-2020)

Lee Doucet

April 21st 2021

1 Abstract

keywords, abstract, introduction, data, model, results, discussion, In the discussion, the paper must include subsections on weaknesses and next steps - but these must be in proportion.

2 Introduction

The topic of preventing school closures is one of those rare bi-partisan issues that can bring people together from both sides of the political spectrum in working towards a common goal. Despite having a shared vision of having access high-quality schools in each community, the topic of school closures are a battleground fought by parents, the school boards, and the provincial government who ultimately decides the rules. This in part stems from the Ministry of Education facing increasing costs while operating schools that are underutilized (Geraghty (2017)). Exacerbating this problem is where the closure occurs as schools have a different function depending on their geographic location. In rural communities, the loss of a school can often lead to negative irreversible impacts as they "are an essential element of the fabric of rural communities" (Geraghty 2017). The loss of any school is devasting can be devasting for those that rely on it. The difference is in the order of magnitude, in Toronto for instance, a school closure can be remedied by traveling an additional stop on the subway. Whereas in rural communities, the difference can transform from a walk to school to a 2-3-hour return trip daily(Geraghty 2017).

At the helm of school closures are urbanization, changes in demographics, and fiscal constraints that require new reforms to keep up with the pace of change (Johns and MacLellan 2020). All of those put substantial pressure on the number of students enrolled in a school to justify its existence. The target the Ontario government is currently looking at is approximately 500 to 800 students per school, which may work as a one size fits all approach in urban settings but may not function as well in rural areas(Irwin B 2021). Rural schools that head steady populations of 150-300 are now being considered too expensive to operate and inadequate as an institution and would benefit from modernization(Irwin B 2021). Once a school is considered to not have a sufficient population, it enters a process called a Pupil Accommodation Review where a decision is made at the school board level regarding to close it or not. The Ministry of Education has tipped the scales towards new builds and closures instead of retrofitting old schools through increased financial support for the former (Robertson 2014). These new builds are not guaranteed to be in the same district as a school board could decide to bus students to another district. This can lead to a cascading problem where a school is shuttered due to low enrollment that leads to families not moving to an area that lost their local school. Which in turn causes a positive feedback loop that that damages the town economically and hurts the areas school age children.

Investigating school closures has certain challenges as there is a lack of public data that has been made available. Both the Ministry of Education and local school boards do not publish the schools that are under review or provide legacy lists of all schools that have been closed and what has replaced them. Even more absent is the reason why the schools were closed in the first place. Meaning information has to be found

through other means, if it can be found, through searching local news articles or statements from parents' groups that have fought school closures. Working within these obstacles, the goal of this research paper is to gather as much information as I can from public data on Ontario's public-school system that can provide data-driven insights into what influences the probability of a school more likely to close. Preference will be made for insights impacting rural communities but a lack of easily-accessible data will limit the scope of the research.

3 Data

3.1 The Source of the Data

The data used in this project was from licensed from the Ontario government, and published on the Canadian government Open Data catalogue(Ontario 2021). It has school enrollment information from both Ontario elementary (junior kindergarten until grade 8) and secondary (grades 9-12) schools. Each academic school year has its own dataset and starts from 2011-2012 and goes until the most recent 2019-2020 academic school year. For a total of 9 data datasets. Note to the reader, I will be using school year interchangeable with academic school year. The files are available in languages French and English with both in XLSX format and text. It was released on 2021-01-27 and was last updated on 2021-01-28.

3.2 Creating a School Closure Dataset

Currently there are no publicly available datasets that I can find that lists public school closures in Ontario. At best, there are anecdotal accounts of school closing published by newspapers but nothing appropriate for data analysis. As a solution, I found Open Data from the Ontario government that listed the enrollment of every school from 2011-2012 to 2019-2020 school years. By comparing a dataset to the following year in sequence, I could tell by the primary key School ID, which schools were not present and had a high probability of closure. I say probability as there is no guarantee that information is correct and needs to be verified during the data pre-processing phase. A new datast was constructed by using the left outer join clause, which includes all rows on the left table (earlier year) that do not match with any years on the right table (year afterwards). Providing a new table that lists all the schools that did not appear in the following year. This process was repeated 8 times and the results were placed in a new dataset.

Before the dataset could be joined to another one, a new column was added called Status which marked all schools as closed. In the 2019-2020 school year data set, the corresponding was also created with each school on the dataset marked as open.They were then joined together, creating 5245 rows of data. Bringing together 4844 schools that were listed as open and 401 schools that were closed. To increase the granularity of the data, a new column region was created that added a region column which placed each school as either south or north depending on their school boards location in Ontario. Separating both regions was the area of Algonquin Park which Statistics Canada uses as part of their map that delineates northern and southern Canada (Canada 2019). ## 3.3 Data Quality

In reference to the data gathered from the Ontario government for use in the custom dataset, overall, the data is good for its intended purposes but has some limitations that prevent it from being high quality. The data was accurate and reliable but had to be cleaned extensively for errors. There were a few instances of a school disappearing in some datasets, which would led me to believe it had closed only for it to return. A couple schools shared the same primary key, School ID, which meant having to find that information outside of the data set. Then there was a problem when trying to attach School IDs to a different database with many of the School IDs not matching even when all other details matched. Despite those difficulties, I still have strong confidence in the data as it is sufficiently fit for the purpose and discrepancies can easily be investigated through a web search to verify supplemental information.

The other area where the data lacked was in the completeness. It would have helped immensely if the data had a clear indicator of how school the was or at minimum, list the city or municipality. Nowhere present

either was any mention of schools that have closed. That being said, there were overall very few missing fields in a 5,000 which drastically cut down the pre-processing time for the data. The data was also very consistent except for a couple naming conventions that were not uniform. For instance, Roman Catholic schools were sometimes just written as Catholic. The structure of the data made it very easy to implement new columns that logically flowed with how the datasets were initialized. Supporting the data was also the timeliness of it, with its age being fairly recent as it was made available for use not long after the previous academic school year. Volatility of the data was minimal, with very few alterations in the data

3.4 Data Strengths

The biggest strength of the data are the demographic information parameters which provide great descriptive insights and make interoperability between datasets possible. In a few instances when using School ID as my primary key did not work between data sets, there was still School Name, Board Number, and Board Name to verify any discrepancies. Information can also be easily attached, Region was created to identify between Southern and Northern Ontario and only required Board Name to add that information to School Name.

To combine with those demographic details, there are plenty of descriptive characteristics to separate schools apart from each other. Schools can be separated by major categories in school language, school level, and school type. Then the lone quantitative parameter, enrollment, allows us to look at the total amount of students in the school boards and track population for its impact. These combined can accomplish a wide variety of different analysis. Are French schools in the north declining over time or are students shifting from one school board to another. Having enrollment also allows the direct comparison between ideal school numbers from the Ministry of Education and the number of students per school.

3.5 Data Ethics

From a statistical perspective, there are missing information that will skew the results from this paper. As mentioned previously, a left outer join between two academic school year data sets provided the schools that were closed, a right outer join was not completed to get a list of schools that were opening. This omission was due to the scope of the project and making it more manageable during the allotted time available for course work. To prevent this paper from misleading anyone, a few summary statistics will help put some of the data into perspective. In the 2011-2012 academic school year, there were 4,899 schools and 2,042,995 total students. By the last year, 2019-2020, there were 4,844 schools and 2,056,055 total students. As you can see, there was not tremendous difference schools and students, what matters is where it is happening.

There were some abnormalities that were not possible to correct with my level of domain expertise. About 68 schools had 15 or less students present with many listings under 10. Some were plausible with Toronto DSB's Native Learning Centre East having 15 students, I could see and welcome a small learning place for Indigenous learners to focus on learning. Others were more difficult to accept, St Martins Catholic School in Toronto has only 15 students enrolled but when I used Google Maps to examine it, it looked much bigger. This will be something to look into the future, I will round the schools under 10 to 10 as I think it makes the most sense.

Attempting to determining what constitutes a rural school was more challenging than first imagined with the present data. To compensate, I used the boundary of Algonquin Park which is close to the line that Statistics Canada uses for separating Northern and Southern Canada. There may be some differences in opinion but across my research, that was what I found that was most common.

Probably the most important omission from my dataset is the reason for closure. There are several reasons why a school closes and each have a different meaning. A school could be dilapidated and another one built in its place, or two schools can be closed as the school board consolidates them into a new school. I've had to be very careful with my research funding to not overstep what the data tells.

Table 1: Table 1: Number of Open/Closed Schools (2020)

School Level	Status	Total Schools	Proportion of Total Schools (%)
Elementary	Open	3965	0.93
Elementary	Closed	312	0.07
Secondary	Open	876	0.91
Secondary	Closed	87	0.09

Table 2: Table 2: School Breakdown Per Region and Type

School Level	School Type	Region	Total Schools	Split of schools (%)
Elementary	Catholic	North	176	0.12
Elementary	Catholic	South	1296	0.88
Elementary	Public	North	249	0.09
Elementary	Public	South	2556	0.91
Secondary	Catholic	North	34	0.12
Secondary	Catholic	South	253	0.88
Secondary	Public	North	82	0.12
Secondary	Public	South	594	0.88

3.5 Descriptive Analysis

3.5.1 Overview of Public School Landscape

From Table 1 we can get a glimpse of the breakdown of the public-school system in Ontario. The vast majority of schools in Ontario are elementary (grades jk-8) with 3,965, compared to 876 secondary schools (grades 9-12). Schools are taught in both the official languages with 4,362 English schools and 479 French schools. Then there is the breakdown of Catholic and Public schools which is more balanced but still favouring public schools with 3,191 public and Catholic Schools. Both of these school levels have eliminated 10% of their schools over a the 8-year period.

Table 1 was generated by Kable(Zhu 2020) & R (R Core Team 2020)

```
## `summarise()` has grouped output by 'School Level'. You can override using the `.groups` argument.
```

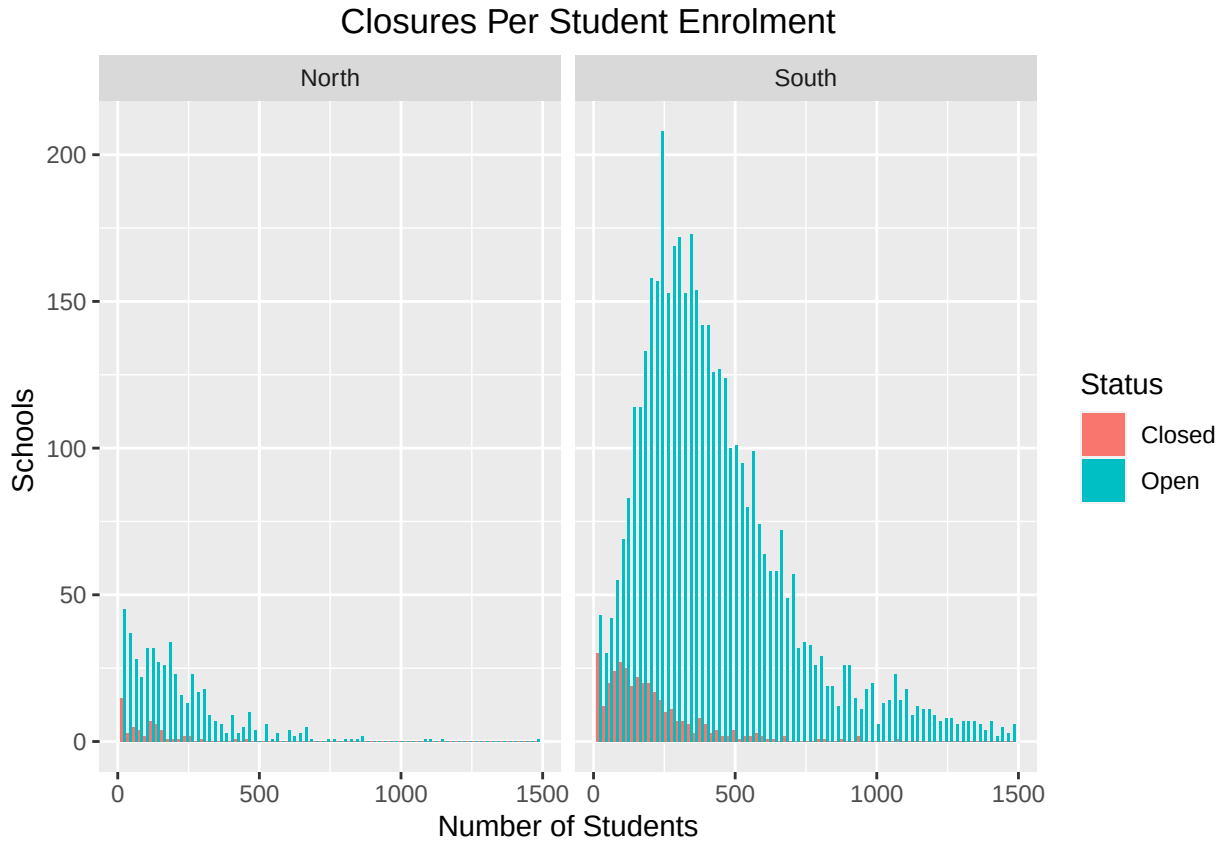
3.5.2 Additional Demographic Paramaters

When we look at some more key demographic information, School Type and Region, we can see in the data two clear important details. First, that the majority of the schools are in the south region. Secondly that Public schools are approximately twice the size of Catholic schools in number. Both Catholic and Public schools have a similar level of school split, except for public schools in the North while have a lower percentage. They may represent evidence of changing school priorities in smaller areas, which the north is as a smaller region.
Table 2 was generated by Kable(Zhu 2020) & R (R Core Team 2020)

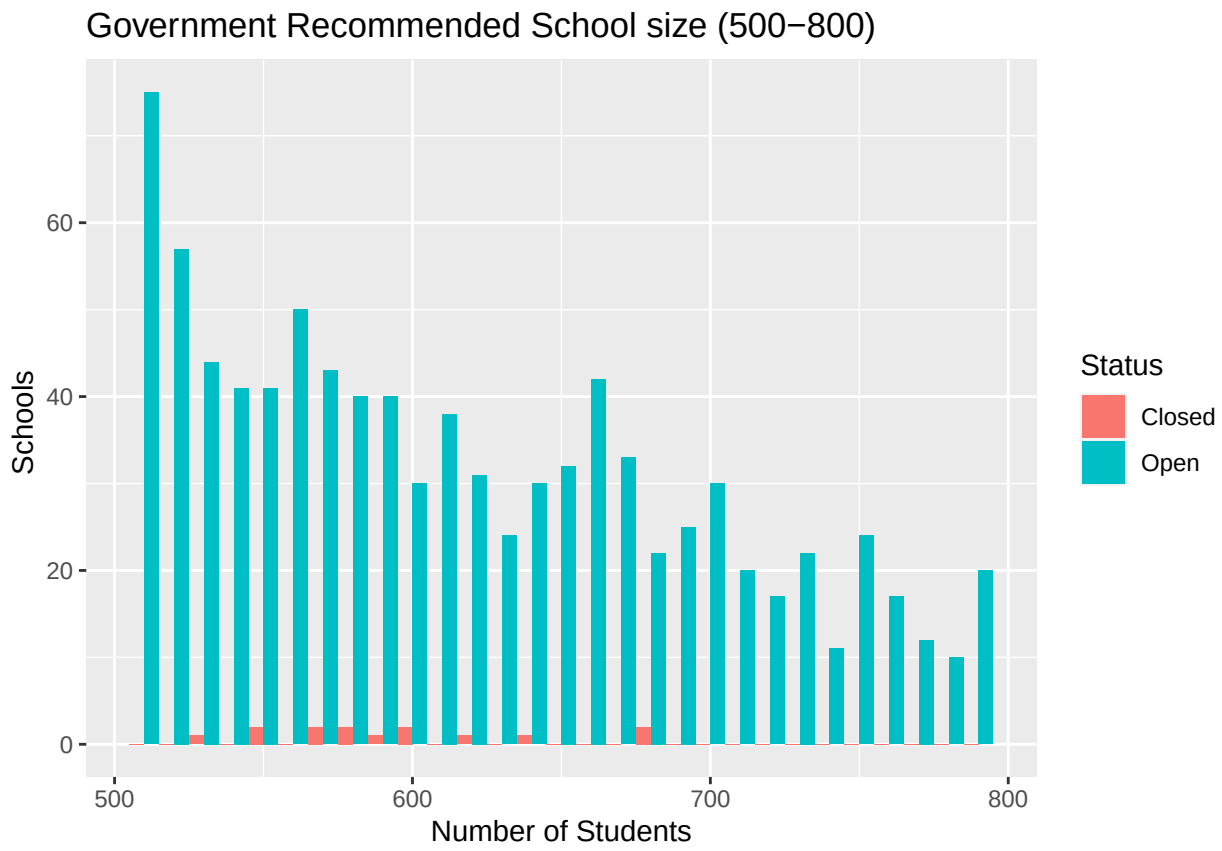
```
## `summarise()` has grouped output by 'School Level', 'School Type'. You can override using the `.group
```

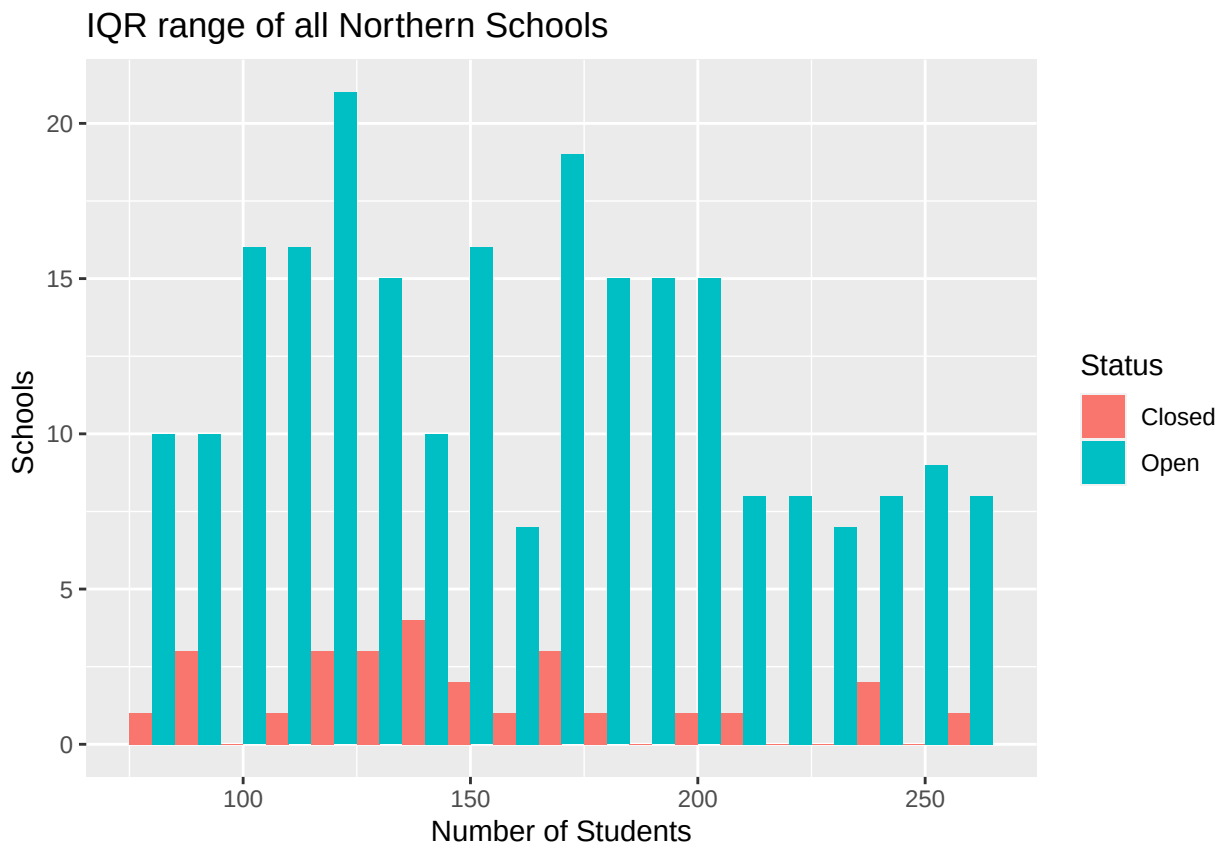
3.5.3 Indications of Closures

Graph 1 highlights just how small northern Ontario school boards are in number compared to their southern counterpart. The graph only lists schools with less than 1500 students for clarity purposes, as onnly a small number of schools proportionately are above that in Ontario. It also provides a good look at the range of students. Southern schools' distribution is similar to a bell curve with the bulk of the students between 230 and 500 students. Whereas Northern schools is more of a slope that gets smaller the more students are enrolled in a school. They both share the same characteristic where the majority of the school closures are closer towards zero. This backs up the research that suggests that underpopulated schools are more likely to close. More surprisingly is that several schools have what I am assuming would be critically low populations. As schools have reported 15 or less students with some reporting as low as 10 students. At this time, it is unknown whether those schools are in the process of closing, represent a clerical error, or have some additional explanation.



Graph 1 was generated by ggplot (Wickham 2016) & R (R Core Team 2020)


Government Recommended School size (500−800)
Schools
60
40
20
0
500
600
700
800
Number of Students
Status
Closed
Open



Graph 2 was generated by ggplot (Wickham 2016) & R (R Core Team 2020) Graph 3 was generated by ggplot (Wickham 2016) & R (R Core Team 2020)

3.5.3 Closures Depending on Enrolment Numbers

Graph 2 shows the ideal range of 500-800 students per school.. If we use the Interquartile range (IQR), it effectively divides up a set of numbers into quartiles (25%, 50%, 75%) which are break downs of the distribution of the data at certain levels. It is useful as it provides snapshots of where the data is and can be used for comparison. When we take graph 1 and examine the range of student enrollment it creates some concerns when we compare the IQR to the ideal range of students. Overall, on average, a school will only have 325 students, much less than the lower bound cut-off of 500 recommended students. Only when we reach the around the 75% of the highest enrollment of schools do they cross into the acceptable threshold. There would have to be dramatic changes for schools to suddenly meet those targets as they are unrealistic at the present

Then we have graph 3 which displays the IQR of schools that are in the north. Unfortunately the top-bound of the IQR is just over half the minimal threshold for desired amount of students per school. This puts schools in the north at a significant disadvantage. You can see it in the numbers when comparing school closures with schools in the ideal range. There were 921 schools listed in the 500-800 range, with 14 of them closing. Then when you compare that to the schools in the north, there were substantially less schools with 233 and almost double the amount of closures with 27.

4 Model

4.1 Sample Strategy for the Model

When I compare the amount of schools closed from 2012-2020, 399, it's a drastically smaller number than the 4841 schools that remain open in 2020. Any comparison in this state would introduce too much bias towards schools that were open and not give meaningful results. To correct for this, I will be taking a subset of the school open data to match the number of schools that were closed. In deciding which schools are to be selected, I will be using simple random sample without replacement (SRSWOR). This means that each individual sample in the schools' that are open data set will have the same opportunity of being selected. The order that they are picked in is not important. The without replacement means that after a school has been selected for the sample, they are not placed back for another chance of becoming re-sampled.

$$N!/n!(N-n)!$$

In figure 1 above, N represents the number of schools in the population, n is the number of sampled schools. The ! represents the factorial component where N! would be N times N-1, times N-2, until completing down until reaching 1.(**???**)

4.2 Statistical Method

In order to understand what parameters can influence whether a school remains open or closed, a statistical test will be required. As the dependent variable is binary, conducting a logistic regression is an appropriate test. Since we are using a subset of the open data, there is no danger in overfitting for that data which is when a model's data is limited and can only speak for internal validity and not external validity.

4.2 Control Variables

The first control variable is enrolment. From looking at the graphs with trends in student's population, it is clear that enrolment is a key indicator of whether a school becomes closed or not judging by the clustering of school closures near overall low student populations.

The second control variable is Region. As part of the main investigation into School Closures, it was important to see if schools belonging to different regions had different impacts

Third was Board Name, this is another way of tackling region at a more granular level.

4.2 Logistic Regressions

4.2.1 Enrolment on Status

Enrolment was very significant with a p-value of <2e-16

4.2.2 Region on Status

No significant findings

4.2.3 Enrolment and Region on Status

Enrolment was very significant with a p-value of <2e-16

4.2.4 Board Name on Status

Out of 75 boards, York Region DSB was the only one under the alpha level of 0.5 with a p-value of 0.0120

4.2.5 Board Name and Region on Status

No significant findings

4.2.6 Board Name and Enrolment on Status

Enrolment was very significant with a p-value of <2e-16. Three Boards had strong effects, Hamilton-Wentworth CDSB with a p-value of 0.00408 and Niagara DSB with a p-value of 0.00834

4.2.7 Board Name, Region, Enrolment, School Level, School Language, and School Type on Status

Enrolment was very significant with a p-value of <2e-16. Secondary Schools' almost had a significant impact with a p-value of 9.91e-09

4.3 Model Prediction

```
inTrain <- createDataPartition(y = sampleSize$Status, p = .60, list = FALSE)
training <- sampleSize[inTrain,]
testing <- sampleSize[-inTrain,]

dim(training)
```

```
## [1] 480  10
```

```
dim(testing)
```

```
## [1] 318  10
```

```
closedProb <- predict(logRegression2, testing, type ="response")
closedPred <- rep("Closed", dim(training)[1])
closedPred[closedProb > .5] = "Open"
table(closedPred, training$Status)
```

```
##
## closedPred Closed Open
##     Closed     31   30
##     Open      209  210
```

```
levels(mydf$Status)
```

```
## [1] "Closed" "Open"
```

```r
mean(closedPred == training$Status)
```

```
## [1] 0.5020833
```

```r
predict(logRegression2, newdata=data.frame(Enrolment=c(525),Region=c("South")), type="response")
```

```
##    1
## 0.51
```

5 Results and Discussion

Appendix

References

Canada, Government of. 2019. "Delineating Northern and Southern Canada." 2019. https://www150.statcan.gc.ca/n1/daily-quotidien/190704/mc-a001-eng.htm.

Geraghty, Shannon. 2017. "The Rural Effect on School Closures and the Limitations Within the Provincial Policy Framework."

Irwin B, Rappolt R, Seasons M. 2021. "School Closures in the Modern Era, the Hollowing Out of Rural Canada." 2021. https://open.toronto.ca/dataset/apartment-building-evaluation/.

Johns, Carolyn, and Duncan MacLellan. 2020. "Public Administration in the Cross-Hairs of Evidence-Based Policy and Authentic Engagement: School Closures in Ontario." *Canadian Public Administration* 63 (1): 117–39.

Ontario, Government of. 2021. "Ontario Public Schools Enrolment." 2021. https://open.canada.ca/data/en/dataset/d89271cf-c5b7-4537-9d8b-5905766d93c6.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robertson, Sean. 2014. "Declining Enrolment in Ontario: What Can History Tell Us and Where Do We Go from Here?" *Canadian Journal of Educational Administration and Policy*, no. 164.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.