# Introduction

The topic of this research is to focus on understanding how students are performing from a collected dataset. By performing exploratory data analysis. The researcher would have the ability to analyze and understand the performance statistically.

Students' performance is critical in many ways. To school and teachers, understanding their performance will help the instructors adjust and pace of teaching and related course materials. To parents, understanding students' performance enables them to keep track of how their kids are doing in school. The performance of students is also important to the public. The public wants to understand what schools would have better infrastructure in supporting students' learning progress and a better chance of improving their performance.

# Analysis and Models

## Reading in the data

The sample data is a rather small dataset. The data contains the progress of students taking the same math course this semester. There are five schools (A, B, C, D, and E). The semester is about ¾ of the way through.

There are 8 variables in the dataset. The variable names and sample data are presented below. The variable school represents where the student took the course. The variable section represents which section of the student was in. Variable very ahead, middling, behind, more behind, very behind, and completed represent how students are performing.

## $ School : chr "A" "A" "A" "A" ...  
## $ Section : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Very Ahead +5 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Middling +0 : num 5 8 9 14 9 7 19 3 6 13 ...  
## $ Behind -1-5 : num 54 40 35 44 42 29 22 37 29 40 ...  
## $ More Behind -6-10: num 3 10 12 5 2 3 5 11 8 5 ...  
## $ Very Behind -11 : num 9 16 13 12 24 10 14 18 12 5 ...  
## $ Completed : num 10 6 11 10 8 9 19 5 10 20 ...

The performance variables were recorded in below fashion.

* very ahead (more than 5 lessons ahead)
* middling (5 lessons ahead to 0 lessons ahead)
* behind (1 to 5 lessons behind)
* more behind (6 to 10 lessons behind)
* very behind (more than 10 lessons behind)
* completed (finished with the course)

In the data cleaning section, it's critical to understand how the researcher identifies the data types of each variable. The researcher defines variable school and section to be a factor. In other words, those two variables are categorical variables. Categorical variables help us label the data. Other variables such as very ahead, middling, behind, more behind, very behind, and completed represent the discrete count of the students in each category. The data should all be integers.

One of the important processes is to identify the null value or missing data. By running the command, we can see there isn’t any missing data. The dataset is very clean.

## Descriptive Analysis

The research of this dataset will be mainly focused on descriptive analysis. The researcher will compose many plots and graphs to understand the data descriptively.

This bar chart represents the number of sections for each school. By reading the plot, we understand A has the most sections. The ranking is A, B, C, D, and E. E has the least sections.

A close up of a logo

Description automatically generated

This bar plot represents the numbers of students who completed the courses on time and which section they are from. The researcher can’t identify any types of distribution in this graph. However, section 30 seems to have more students who completed the courses on time followed by 10, 27, and 7.

A screenshot of a cell phone

Description automatically generated

This graph presents the number of students who are very ahead comparing to other students. The plot shows there aren't any students who are very ahead in the course.

A close up of a tree

Description automatically generated

This bar plot represents how many students are in the middle status. Section 7 takes the lead followed by 4 and 10.

A close up of a device

Description automatically generated

This plot represents the number of students who are behind. Section 30 has a number of students who are behind.

A screenshot of a cell phone

Description automatically generated

Below two plots represented how many students are more and very behind in the course. Section 2 to section 14 has a higher number of students who are more behind and very behind.

A close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generated

We can also review the boxplots for each progress category. We are able to quickly find some outliers in completed and more behind categories. Behind category also has a wider range of students.

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated | A screenshot of a social media post  Description automatically generated |
| A screenshot of a cell phone  Description automatically generated | A screenshot of a cell phone  Description automatically generated |
| A screenshot of a social media post  Description automatically generated |  |

Further investigate progress completed and more behind. We can find school E and section 1 has the highest number of students (27) who completed the course. We can also find school A has more students who are in “more behind” progress status.

Here is another way of finding outliers – log transformation. We can easily identify the outliers in each category. It also shows that there are more data points in behind category.

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated | A screenshot of a cell phone  Description automatically generated |
| A screenshot of a cell phone  Description automatically generated | A screenshot of a cell phone  Description automatically generated |
| A screenshot of a cell phone  Description automatically generated |  |

This plot shows that most of the students are behind in the course. Also, the statistics show that 20% of students completed the course, 0% of students are very ahead, 14% of students are middling, 47% of students are behind, 6% of students are more behind, and 13% of students are very behind. That being said, there are almost 70% of students who are behind in the course.

A close up of a logo

Description automatically generated

## Completed Very Ahead +5 Middling +0 Behind -1-5   
## 0.19737664 0.00000000 0.13866334 0.47095565   
## More Behind -6-10 Very Behind -11   
## 0.06246096 0.13054341

We can also review the correlation between completed and other categories. We don’t really see any strong correlations between those categories by the plots.

|  |  |
| --- | --- |
| A close up of text on a white background  Description automatically generated | A screenshot of a cell phone  Description automatically generated |
| A close up of text on a white background  Description automatically generated | A screenshot of a cell phone  Description automatically generated |

It’s also not significant when we generate a matrix of correlations.

## Completed Very Ahead +5 Middling +0 Behind -1-5  
## Completed 1.0000000 NA 0.4166307 0.2443381  
## Very Ahead +5 NA 1 NA NA  
## Middling +0 0.4166307 NA 1.0000000 0.4078917  
## Behind -1-5 0.2443381 NA 0.4078917 1.0000000  
## More Behind -6-10 0.1323454 NA 0.2095337 0.5589297  
## More Behind -6-10  
## Completed 0.1323454  
## Very Ahead +5 NA  
## Middling +0 0.2095337  
## Behind -1-5 0.5589297  
## More Behind -6-10 1.0000000

Now, we would think the course must be very challenging that most of the students are behind. Is that true? How are the students performing in each school? The below plots show the counts of students' progress for each school.

|  |  |
| --- | --- |
| A close up of a logo  Description automatically generated | A close up of a logo  Description automatically generated |
| A close up of a logo  Description automatically generated | A close up of a logo  Description automatically generated |
| A close up of a logo  Description automatically generated |  |

It’s understood that most of the students are behind even in the breakdown of each school. However, we do notice that school B has more students who completed the course.

This plot has presented the students who have completed the course from school B by section. We found that section 6, 10, and 12 have the better results within school B.

A screenshot of a cell phone

Description automatically generated

This plot has presented the students who are behind the course from school B by section. We found that section 1, 3, and 9 have higher numbers of being behind within school B.

A close up of a logo

Description automatically generated

# Results

# With exploratory analysis, we have below important findings.

1. The students' learning abilities and performance are not normally distributed across schools and sections.
2. By reviewing the boxplot, we understand that the average number of students is higher in the Behind category.
3. There are not many outliers in students' learning progress. By looking at the log transformation, we can see one or two outliers in completed, middling, behind, and very behind categories.
4. The distribution of learning progresses is very similar across all five schools.
5. Over 47% of students are behind, 6% of students are more behind, and 13% of students are very behind. In other words, there are almost 70% of students who are behind in the course.
6. School B has a greater number of students who completed the course.

# Conclusions

With exploratory data analysis, the researcher can understand how the students are performing in the course. It also gave the researcher an opportunity to explore the data and ask questions such as – which school has students with better performance? By observing the data as a third party, we analyze the data with statistical methods. We found that school B has a higher percentage of students who completed the course. It's worth exploring how school B achieves better performance and if it can be applied to other schools.