Joyce Woznica

jlwoznic@syr.edu

October 18, 2019

Joyce Woznica  
Homework #2

Contents

[Introduction 3](#_Toc22308729)

[Analysis 3](#_Toc22308730)

[The Data 3](#_Toc22308731)

[Data Load, Cleanse, Munge and Preparation 4](#_Toc22308732)

[Exploratory Data Analysis 5](#_Toc22308733)

[Descriptive Statistics 5](#_Toc22308734)

[Visualization 9](#_Toc22308735)

[Results 14](#_Toc22308736)

[Conclusions 15](#_Toc22308737)

List of Figures

[Figure 1: Slate Rock and Gravel partnered with Bedrock High Schools in Bedrock 3](#_Toc22308790)

[Figure 2: Dataset Description 4](#_Toc22308791)

[Figure 3: Before and After Column Names 4](#_Toc22308792)

[Figure 4: Verification of No Missing Data 4](#_Toc22308793)

[Figure 5: Dataset after Cleanse and Preparation 5](#_Toc22308794)

[Figure 6: Summary Statistics for School Data 5](#_Toc22308795)

[Figure 7: Describing the School Data 6](#_Toc22308796)

[Figure 8: Summary Statistics of Lesson Status Variables of School Data 6](#_Toc22308797)

[Figure 9: Summary Students in each Lesson Category by School 7](#_Toc22308798)

[Figure 10: Total Number of Student by Lesson Status and Portion of All Students 7](#_Toc22308799)

[Figure 11: Melted Data Frame with Lesson Status Factors 8](#_Toc22308800)

[Figure 12: Mean Number of Students in each Lesson Status Category 8](#_Toc22308801)

[Figure 13: Mean Number of Students in each Lesson Status Category 9](#_Toc22308802)

[Figure 14: Students in each Lesson Status Highlighted by School 9](#_Toc22308803)

[Figure 15: Students in each Lesson Status Highlighted by School 10](#_Toc22308804)

[Figure 16: Mean Number of Students in each Lesson Category 10](#_Toc22308805)

[Figure 17: Students in each Lesson Status Highlighted by School without Totals 11](#_Toc22308806)

[Figure 18: Students in each Lesson Status grouped by School 11](#_Toc22308807)

[Figure 19: Total Students at each School 12](#_Toc22308808)

[Figure 20: Total Sections by School 12](#_Toc22308809)

[Figure 21: Students in each Lesson Status grouped by Schools B and E 13](#_Toc22308810)

[Figure 22: Students in each Lesson Status grouped by School 14](#_Toc22308811)

# Introduction

There is a current promotion in the Town of Bedrock to implement self-paced study and improve the potential of the upcoming high school students of being accepted into college. This program is funded by a partnership between the Slate Rock and Gravel Company and the Town of Bedrock to help find qualified individuals to intern and eventually employ in new data analytic and scientist positions at the company.

Figure 1: Slate Rock and Gravel partnered with Bedrock High Schools in Bedrock

Part of the new curriculum includes a math course has been started in the schools in town. The course allows students to work independently to move ahead if they are ready despite the overall progress for the class. This allows those with the potential to move ahead while allows others that need more time to work at their own pace. The only criterion is that the entire set of lessons within the semester.

A study is currently being conducted to gage the successfulness of this program and to review the current state at the schools in the town as the semester is moving into its fourth and final quarter. The hope is that any extraneous information can be identified and other trends or issues can be discovered.

This document outlines the data collected, the loading process, any modifications required to the data and why, initial visualizations and information about the data, the results of various modeling techniques and conclusions drawn from the analysis.

# Analysis

## The Data

Each of the five schools (A, B, C, D and E) have implemented the same 35-lesson math course this semester. Spread over the five schools, there are 30 sections of students taking the class. A description of the dataset can be found in the following table.

| **Variable** | **Definition** |
| --- | --- |
| ***School*** | The school name in the Town of Bedrock (A, B, C, D and E) |
| ***Section*** | The section number within the school that is studying this math course |
| ***Completed*** | The number of students in the section that have *completed* the course |
| ***Very Ahead*** | The number of students in the section that are **more than 5 lessons** *ahead* |
| ***Middling*** | The number of students in the section that are **between 0 and 5 lessons** *ahead* |
| ***Behind*** | The number of students in the section that are **between 1 and 5 lessons** *behind* |
| ***More Behind*** | The number of students in the section that are **between 6 and 10 lessons** *behind* |
| ***Very Behind*** | The number of students that are **more than 10 lessons** *behind* |

Figure 2: Dataset Description

The data set, which will be referred to as the “School Data”, has 30 observations spread over these eight fields or variables. Each school could have a different number of sections offering the specific course.

### Data Load, Cleanse, Munge and Preparation

#### Data Load

To load the School Data, the *read.csv* function was used with all empty data replaced with spaces. In addition, the column names, which contained by “.”’s and blank space were manipulated so that the data would be easier to use in R.





Figure 3: Before and After Column Names

#### Data Cleanse

A check was run on this resulting loaded data using *is.na* to determine if any fields were missing data. It was determined by this verification that there was no missing data.



Figure 4: Verification of No Missing Data

#### Data Munge and Preparation

To properly work with the data, the “Section” field was converted from integer to a factor using the *factor* function. In addition, the columns were reordered from “Completed” to “Very Behind” for easier usage. The results of the data load and cleanse can be found in the view of the data as shown in Figure 5.

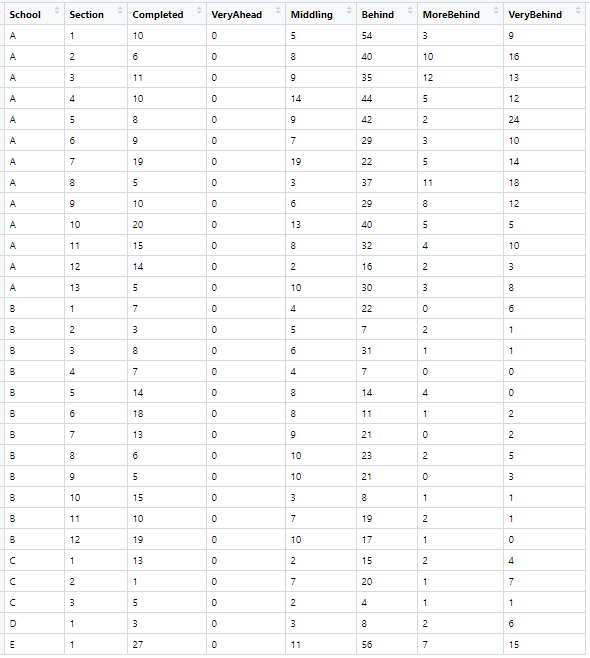


Figure 5: Dataset after Cleanse and Preparation

## Exploratory Data Analysis

### Descriptive Statistics

To gather a general understanding of the data and information in the dataset, some general descriptive statistics were run on the overall data using the functions *summary*, *stat.desc*, *describe* and *stargazer*. The results of these can be found in the following figures.

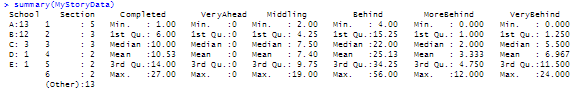


Figure 6: Summary Statistics for School Data

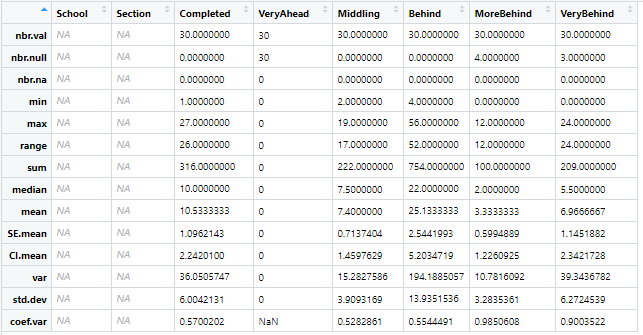
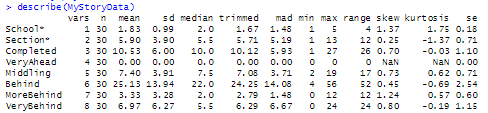
  
  


Figure 7: Describing the School Data

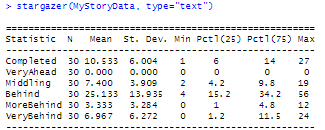


Figure 8: Summary Statistics of Lesson Status Variables of School Data

To gain some further insight into students in each “Lesson Status” category (Very Ahead, Completed, etc.), additional data frames were created with various summary information to assist in reviewing the data. The first was to add all the students in all sections at a particular school within a particular Lesson Status as well as the total number of students at that school. The resulting data frame “SumBySchoolDF” is shown in the following figure.

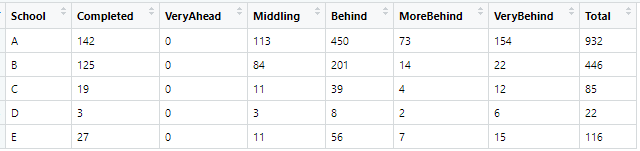


Figure 9: Summary Students in each Lesson Category by School

To gather information about the total number of students in each lesson category, additional manipulation was done to create a smaller data frame with this information as well as the ratio (or percentage) of the total 1,601 students in the study.

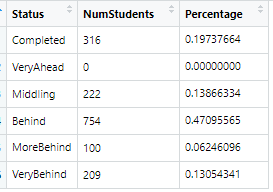


Figure 10: Total Number of Student by Lesson Status and Portion of All Students

This data was then transformed using the *melt* function from the *reshape2* package so that visualizations could later be run on the data. Using this functionality, the Lesson Status became a factor with six levels.

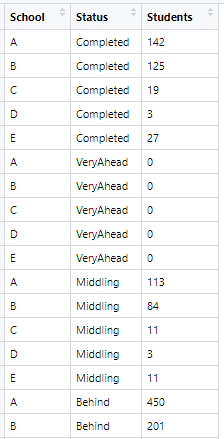
  


Figure 11: Melted Data Frame with Lesson Status Factors

The same melting was done on the full dataset and stored in “meltMyStoryData” for later visualization.

Finally, the mean number of students for each Lesson Status category was created to allow clear information on where all the students regardless of the school or section attending fall in the program.



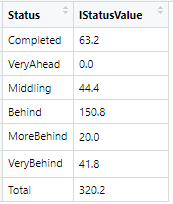


Figure 12: Mean Number of Students in each Lesson Status Category

### Visualization

After reviewing the initial statistics and other information gathered with the descriptive statistics, general plots were completed to visualize the data in different ways for clearer understanding. First a box plot was created by Lesson Status which clearly indicates that the majority of students are in the “Behind” category which means they are one to five lessons behind in the curriculum.

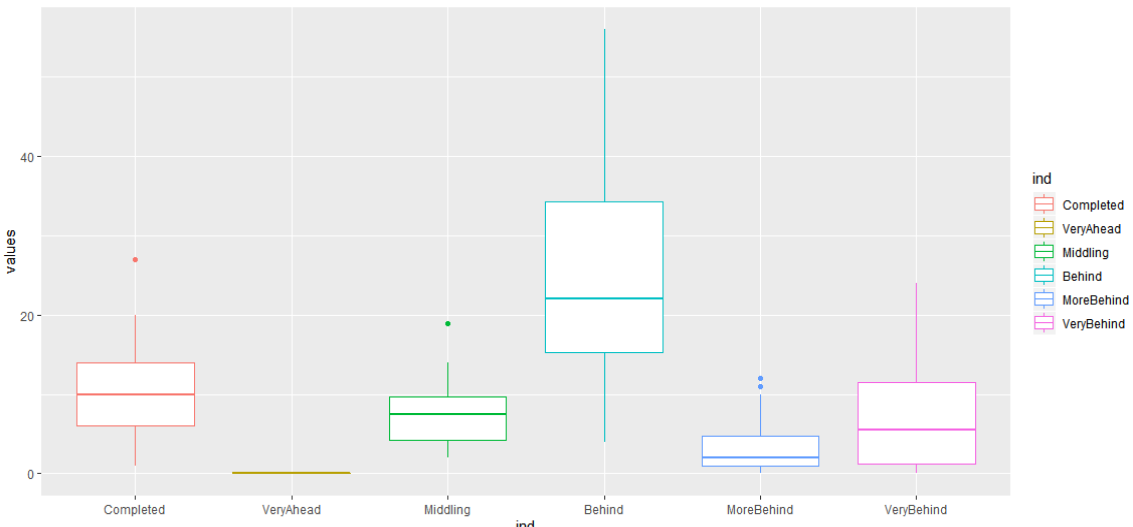


Figure 13: Mean Number of Students in each Lesson Status Category

Based on this information, the same information was plotted using a bar chart with each school highlighted in a different color showing all students in each lesson category as displayed in the following figure.

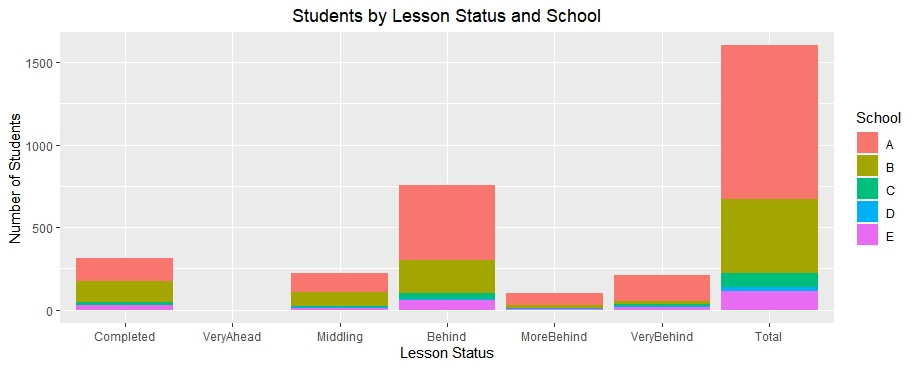


Figure 14: Students in each Lesson Status Highlighted by School

The same information was graphed a bit differently to show the details in a different visualization as reflected in Figure 15.

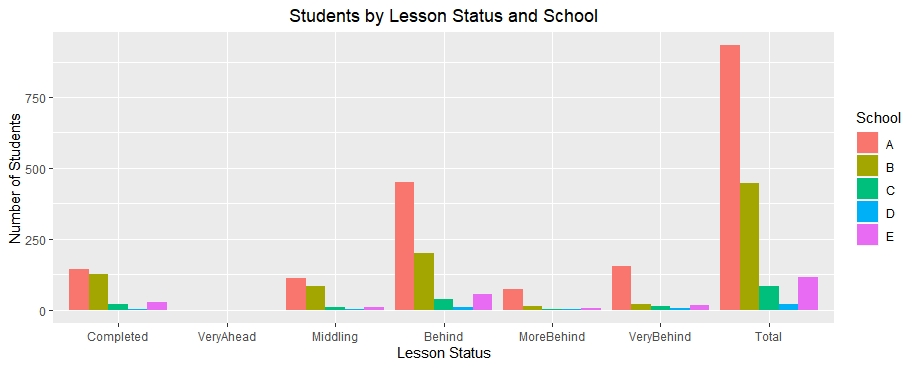


Figure 15: Students in each Lesson Status Highlighted by School

To verify this information, the mean for each status for all schools combined was plotted again showing the larger group in Behind, but also reflecting that there is a significant number of students that have also Completed the curriculum as shown in the figure below.

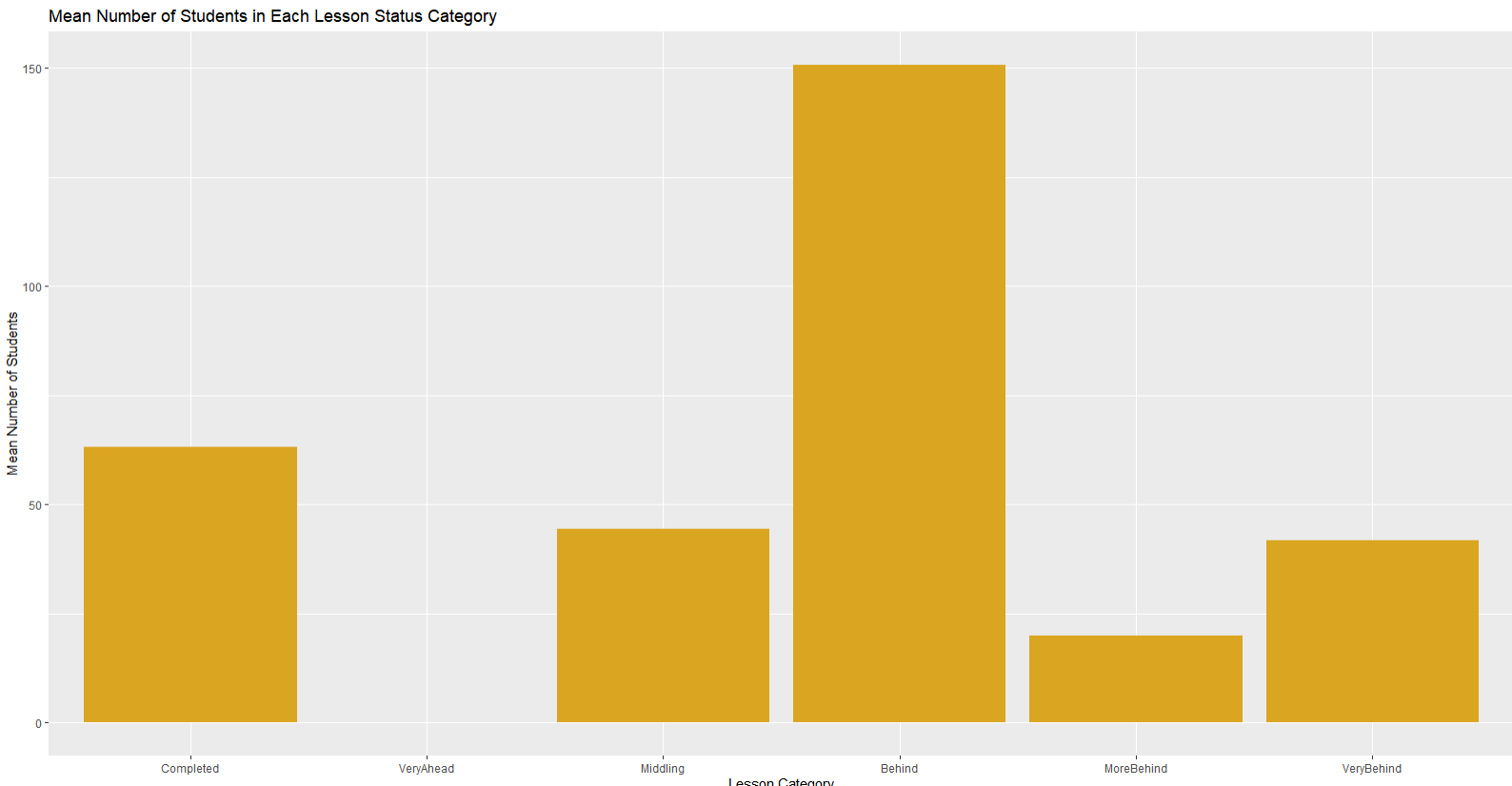


Figure 16: Mean Number of Students in each Lesson Category

This information proved very interesting, so the data was plotted again but without the totals to give a better view of the number of students in each lesson category as detailed in the following figure. This exemplifies the number of individuals that appear in the Behind status.

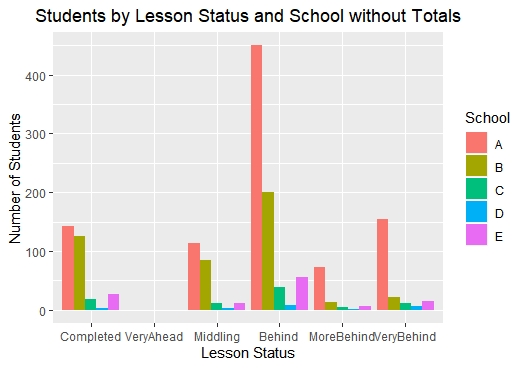


Figure 17: Students in each Lesson Status Highlighted by School without Totals

To take a closer look at this information, the same information was plotted with all statuses grouped by school to determine if any pattern is consistent in each school as shown in Figure 18.

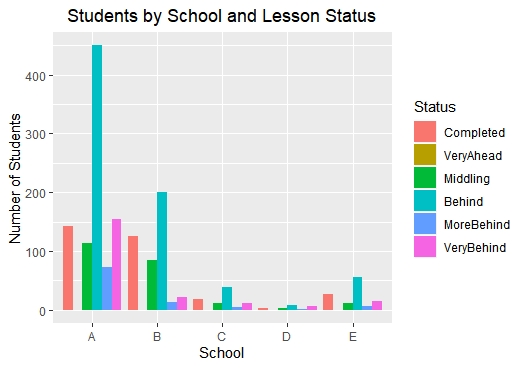


Figure 18: Students in each Lesson Status grouped by School

As can be seen, there are more student falling in the Behind area than any other category, but you can also see in the above plot that the number of students that have Completed the course is much higher at School B and School E than at the other schools.

To put things more in perspective, it was important to see the total number of students and sections at each school since School A has a much larger group of students in the Behind category. These plots are represented in Figure 19 and Figure 20.

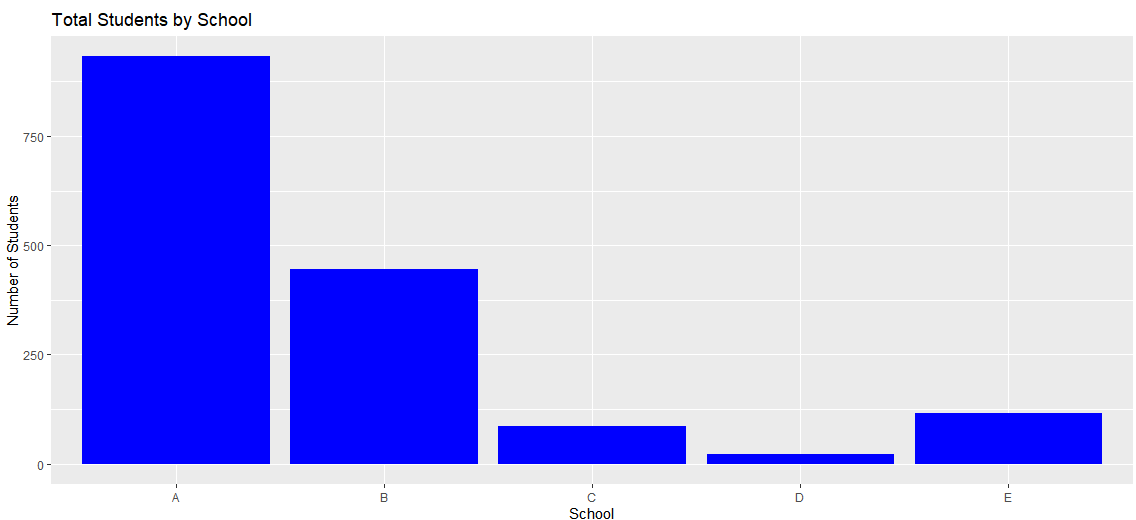


Figure 19: Total Students at each School

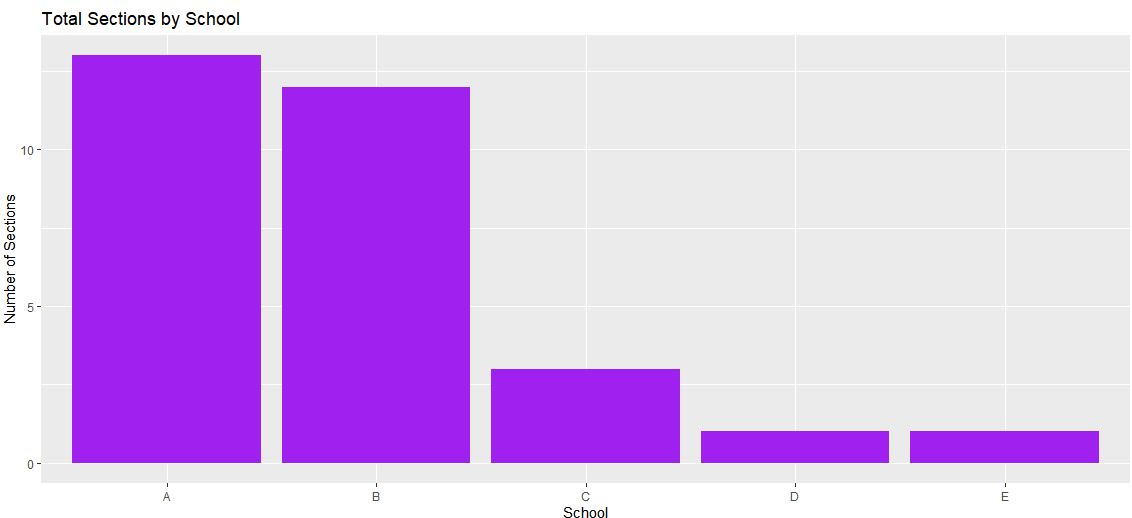


Figure 20: Total Sections by School

Although E had only one section of the math course, this school had far more complete it than other schools in relationship to the total number of students that took the course. The only school that showed similar was B. This is reflected in the plot presented in Figure 21.

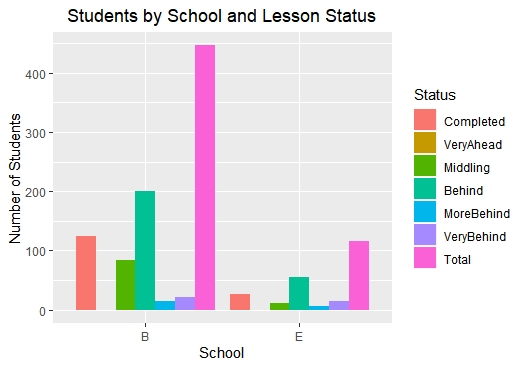


Figure 21: Students in each Lesson Status grouped by Schools B and E

Finally, a scatterplot taking advantage of color and shapes was created to stack the schools against each other to see if any further conclusions could be determined. This is reflected in the following figure.

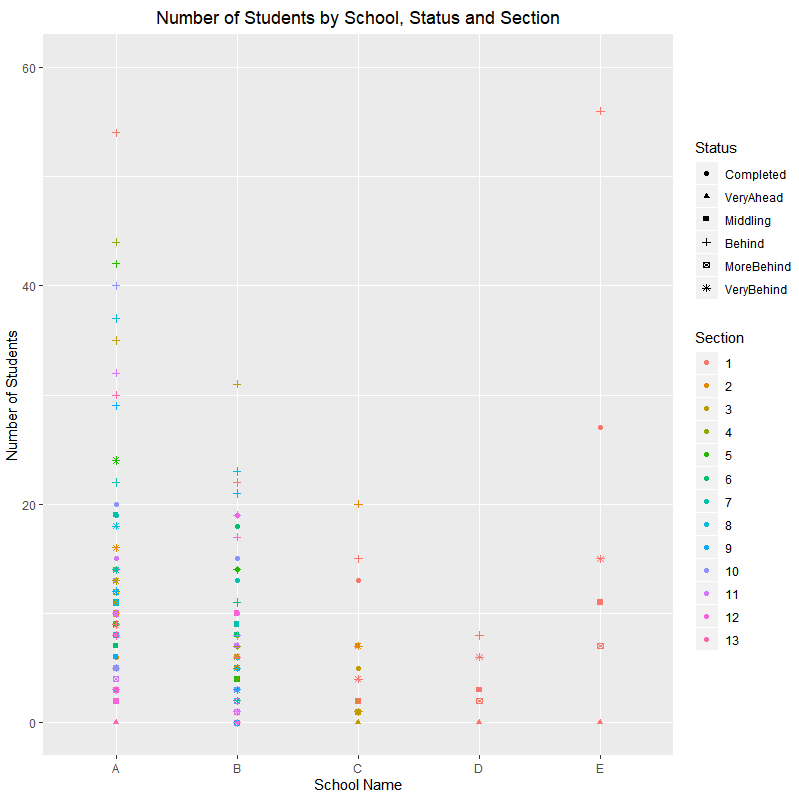


Figure 22: Students in each Lesson Status grouped by School

# Results

As shown in the previous section, a few patterns or lack of patterns were determined based on the descriptive statistics and visualizations presented. These have been summarized in this section.

* It is clear that the majority of students in the data provided fall into the “Behind” category.
* No students are in the “Very Ahead” grouping.
* Although many schools have a similar pattern or percentage of students in each category, schools B and E have a larger percentage of students that have completed the course.

However, school E has very few students overall and this makes it hard to assume that it is a good representation of the general results.

# Conclusions

As a result of the analysis done on this data, the following conclusions were drawn:

* Find out what School B is doing differently, if anything.

Since School B represents a significant set of students and have seen more students complete the program, they may be doing something different and this would be important to understand.

* Research why so many students are behind.

This information can be misleading, since Behind represents students that are between one and 5 lessons behind. One lesson behind is different than five, so it is possible that a large percentage of students in this category are merely one lesson behind.

However, most importantly is that this data is missing information that could help be used to determine what is really happening in these schools.

* For example, grades for the students would be useful as it may be determined that all that have completed are actually the best students and that accounts for much of the success of the program in those schools.
* Teacher information would also be useful. Are the teachers known to be successful? Have they taught this type of curriculum before?
* Demographics are also important. Where the school is located and the general success of the students overall.

Because of this missing information, it is difficult to draw many conclusions, except that School B has figured something out and should be used as a model if this is the only data we can collect about the course.