2019-0703 IST 707 Data Analytics

Homework Assignment 1 (week 1)

Multiple Schools - Math Course, Lessons Progress Reports

Ryan Timbrook

NetID: RTIMBROO
Course: IST 707 Data Analytics
Term: Summer, 2019
Due Date: 7/17/19

Homework Assignment 1 (week 1)

Table of Contents

1 Introduction

Five schools (A, B, C, D and E) have implemented the same math course, with 35 lessons, for the current school term. The semester is approaching the 75%-waymark. The school district would like to get a sense of how each school is progressing in completing all 35 lessons and if there are any early warning indicators that could be understood.

For each section of this course, collected over the five schools, records were taken, counting the number of students who were in one of six groupings: 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) and Completed (finished with the course)

Based on the data, and time left in the term, are there indicators that can show how one school is performing over another? What stories does the data reveal?

2 Analysis and Models

2.1 About the Data

The data set contains 30 records of 8 variables. It's made up of one Factor variable, the School, and seven numeric continuous variables, six of which are the counts of students grouped into a lessons progress categorization. The six lesson progress categorizations are: Very.Ahead..5, Middling..0, Behind..1.5, More.Behind..6.10, Very.Behind..11, Completed. The Section numeric variable represents a school's section number for this Math course. Each school, due to its student class size has a different number of sections.

There were no missing values from the original dataset. In the initial cleaning, some column headings were altered for readability purposes.

Original Column Name	New Column Name
Very.Ahead..5	VeryAhead
Middling..0	Middling
Behind..1.5	Behind
More.Behind..6.10	MoreBehind
Very.Behind..11	VeryBehind

2.1.1 *Dataset Info*

Each of 5 schools (A, B, C, D and E) have the same math course for the current semester, with 35 lessons. There are 30 sections total.

Each section has a count of students who are:

- 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)

Attribute	Data Type	Preview
School	Factor w/5 levels	'A','B','C','D','E',…
Section	Int	1 2 3 4 5 6 7 8 9 10 …
VeryAhead	Int	0 0 0 0 0 0 0 0 0 0 …
Middling	Int	5 8 9 14 9 7 19 3 6 13 …
Behind	Int	54 40 35 44 42 29 22 …
MoreBehind	Int	3 10 12 5 2 3 5 11 8 5 …
VeryBehind	Int	9 16 13 12 24 10 14 18 …
Completed	int	10 6 11 10 8 9 19 5 10 …

2.1.2 *Data Exploration & Cleaning*

- View of the head of the dataset to get a sense of how the data looks.

Table 2.0: Preview (head) of Schools Dataset

School	Section	VeryAhead	Middling	Behind	MoreBehind	VeryBehind	Completed
A	1	0	5	54	3	9	10
A	2	0	8	40	10	16	6
A	3	0	9	35	12	13	11
A	4	0	14	44	5	12	10
A	5	0	9	42	2	24	8
A	6	0	7	29	3	10	9

- View of the summary statistics of the school dataset as it was obtained.

The school variable shows a disproportionate number of records per school. This represents the number of math course sections in a given school. It's reflective of the overall number of students proportional to the schools. (i.e., More students in a school, more sections their will be.) The VeryAhead attribute is 0 for all schools, this attribute can be removed from most analysis. Most of the means and medians are close representing a normal distribution, except for 'Behind', where it shows a higher mean value, skewing the dataset distribution to the right.

Table 2.1: Summary Table of Schools Dataset

School	Section	VeryAhead	Middling	Behind	MoreBehind	VeryBehind	Completed
A:13	Min. : 1.00	Min. :0	Min. : 2.00	Min. : 4.00	Min. : 0.000	Min. : 0.000	Min. : 1.00
B:12	1st Qu.: 2.25	1st Qu.:0	1st Qu.: 4.25	1st Qu.:15.25	1st Qu.: 1.000	1st Qu.: 1.250	1st Qu.: 6.00
C: 3	Median : 5.50	Median :0	Median : 7.50	Median :22.00	Median : 2.000	Median : 5.500	Median :10.00
D: 1	Mean : 5.90	Mean :0	Mean : 7.40	Mean :25.13	Mean : 3.333	Mean : 6.967	Mean :10.53
E: 1	3rd Qu.: 9.00	3rd Qu.:0	3rd Qu.: 9.75	3rd Qu.:34.25	3rd Qu.: 4.750	3rd Qu.:11.500	3rd Qu.:14.00
NA	Max. :13.00	Max. :0	Max. :19.00	Max. :56.00	Max. :12.000	Max. :24.000	Max. :27.00

- Describe of the data

Table 2.2: Describe Table of School Dataset

vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	
School*	1	30	1.833333333	0.985527457	2	1.666666667	1.4826	1	5	4	1.365855503	1.750089866	0.179931873
Section	2	30	5.9	3.898275481	5.5	5.708333333	5.1891	1	13	12	0.248276745	-1.373129235	0.711724472
VeryAhead	3	30	0	0	0	0	0	0	0	0	NA	NA	0
Middling	4	30	7.4	3.909316899	7.5	7.083333333	3.7065	2	19	17	0.733247846	0.621436528	0.71374035
Behind	5	30	25.13333333	13.9351536	22	24.25	14.0847	4	56	52	0.448271924	-0.693528138	2.544199322
MoreBehind	6	30	3.333333333	3.283536081	2	2.791666667	1.4826	0	12	12	1.244966939	0.568447664	0.599488927
VeryBehind	7	30	6.966666667	6.272453919	5.5	6.291666667	6.6717	0	24	24	0.802207619	-0.194324181	1.145188167
Completed	8	30	10.53333333	6.00421308	10	10.125	5.9304	1	27	26	0.696787674	-0.030093955	1.096214315

- No cleaning Necessary, there were no incomplete data records

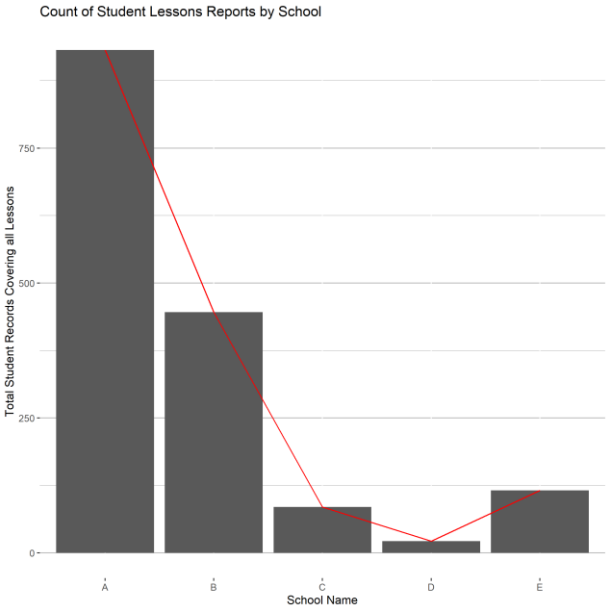
2.1.3 *Data Transformations*

- Dimensionality Reduction:
  - Section attribute was removed from most analysis. It was found to be proportional to the number of students enrolled in the math course for a given school. Each of the schools having a disproportional number of students to one another.
  - VeryAhead attribute was found to be 0 for all sections of each school. It was removed from most analysis and visualizations.
- Aggregation: Total student counts summed by school

Summation of progress groups by school along with adding a new attribute, 'Totals', to the dataset. The graph below, 'Count of Student Lessons Reports by School', visualize the distribution of students across the five schools.

Table 2.3: Total counts of student lessons distributed across schools

School	VeryAhead	Middling	Behind	MoreBehind	VeryBehind	Completed	Totals
A	0	113	450	73	154	142	932
B	0	84	201	14	22	125	446
E	0	11	56	7	15	27	116
C	0	11	39	4	12	19	85
D	0	3	8	2	6	3	22

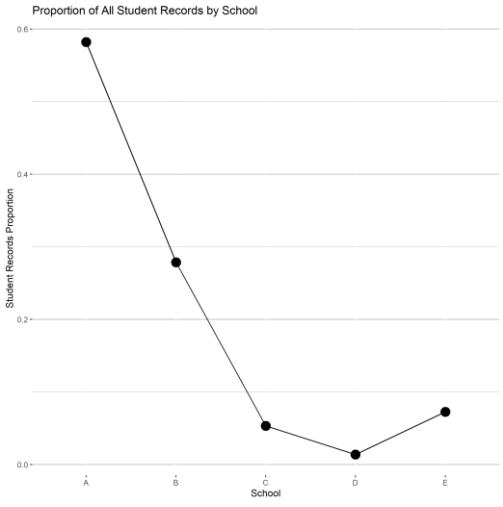


- Normalization: Proportion's of Students at each School to total population

  The table 2.4 below represents the overall proportion of students at each school to the total number all students enrolled in the same math course at each of the five schools. The graph, 'Proportion of All Students Records by School', helps to visualize the school's student populations inequalities.

  Table 2.4: Proportion of students to each School

School	Total.PROP
A	0.5821
B	0.2786
C	0.0531
D	0.0137
E	0.0725

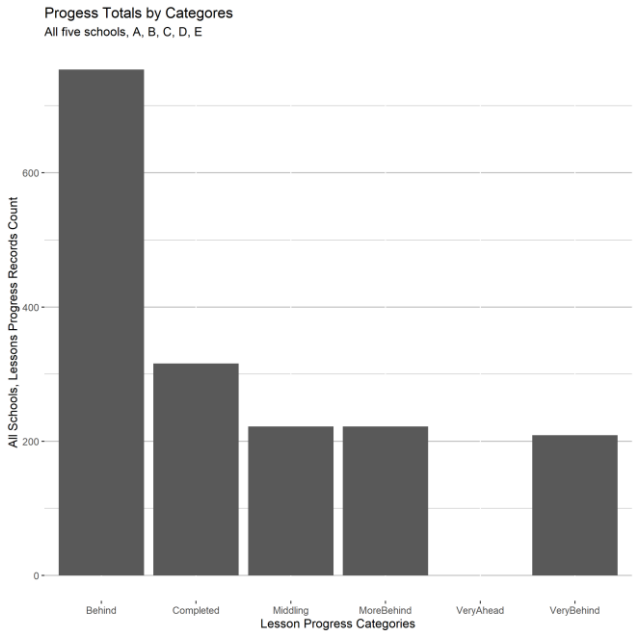


- Aggregation: Total lesson progress counts by grouping of all schools

  Summing each of the lessons progress groupings and visualizing them to represent overall distribution of students in each of these categories shows that 'Behind' has more than twice the number of students than its closest category, Completed. And more than three times 'Middling', 'MoreBehind' and 'VeryBehind'.

  Table 2.5: Progress Categories Counts

Progess Category	Totals
Behind	754
Completed	316
Middling	222
MoreBehind	222
VeryBehind	209
VeryAhead	0



- Normalized Dataset to Proportions of total students by School

  An approach of normalizing each school's student lesson progress distribution was taken in order to compare the school's performance to each other. Each grouping's value is a proportion of students in that group to the total population of the given school. Equally distributing, or ranking, based on the number of students enrolled in the match course at that school.
  Table 2.6 below is a view of the new, proportions based, school dataset. It also has three new attributes, (GREEN, YELLOW, RED), used for alert levels that will be described in another section.

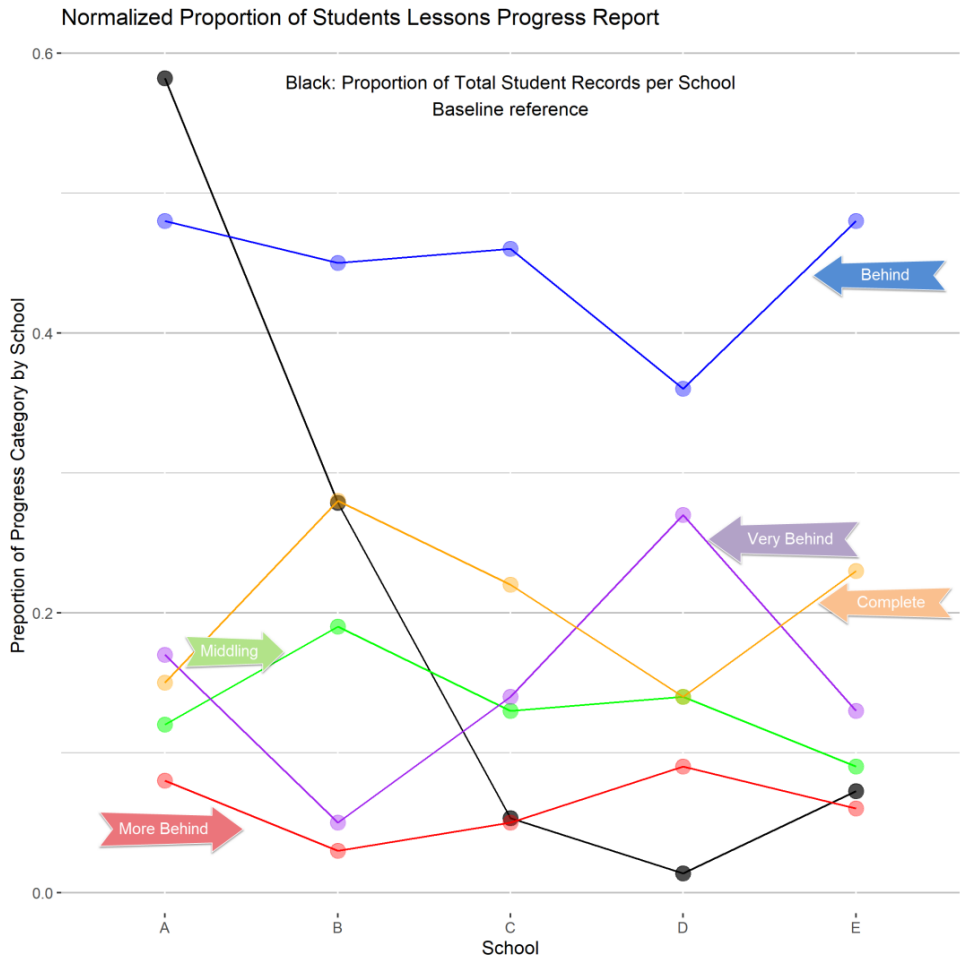
  Shown in Graph 2.6.1, 'Normalized relationships of progress groupings by school', is a layered visualization that shows how each school is performing for a given lessons progress grouping in relation to all other schools.

  In Graph 2.6.2 below, each lesson progress grouping is broken out into its own visualization that incorporates the overall average distribution. This visual helps to show how far or close to the norm a school is for that topic. As an example, for 'VeryBehind', school's B and D show the greatest distance from the average and each other. School B is only 5% VeryBehind, where school D is 27% VeryBehind.

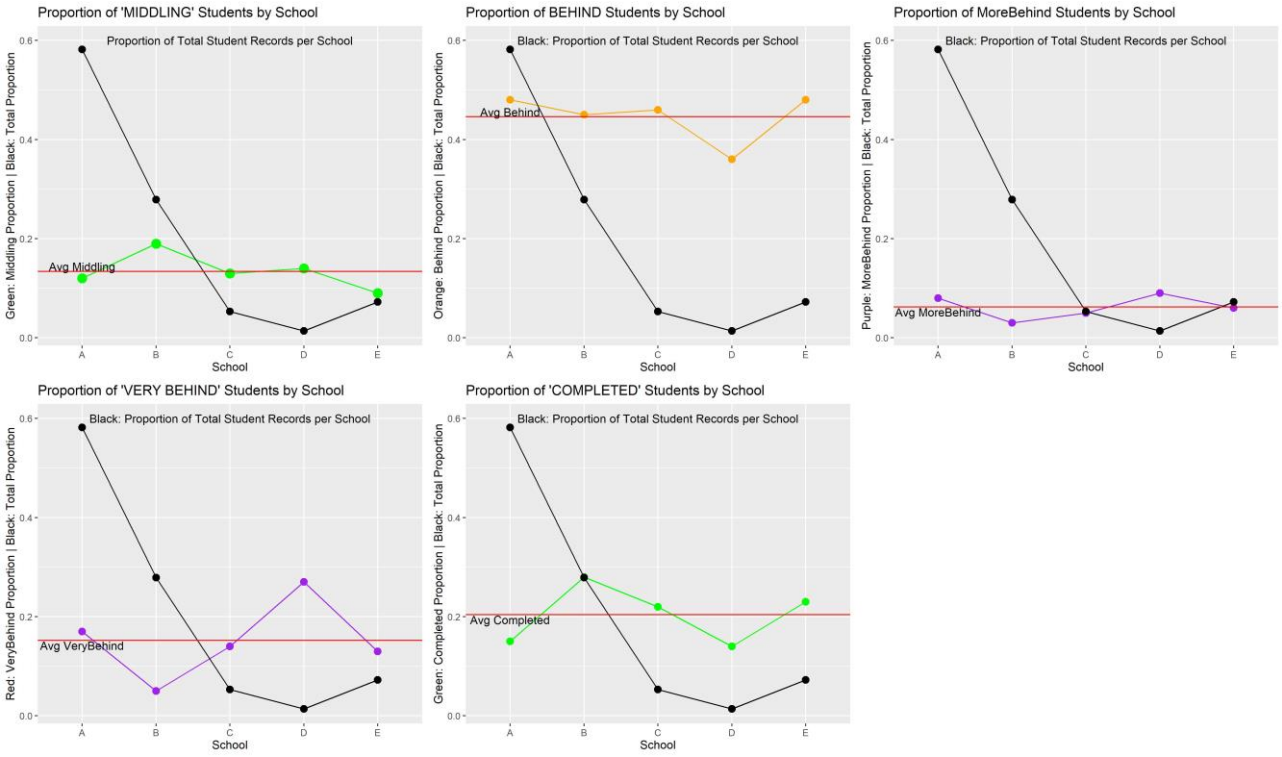
Table 2.6: Table view of school dataset as proportions to student count by school

School	VeryAhead.PROP	Middling.PROP	Behind.PROP	MoreBehind.PROP	VeryBehind.PROP	Completed.PROP	GREEN.PROP	YELLOW.PROP	RED.PROP	Total.PROP
A	0	0.12	0.48	0.08	0.17	0.15	0.27	0.48	0.24	0.5821
B	0	0.19	0.45	0.03	0.05	0.28	0.47	0.45	0.08	0.2786
C	0	0.13	0.46	0.05	0.14	0.22	0.35	0.46	0.19	0.0531
D	0	0.14	0.36	0.09	0.27	0.14	0.27	0.36	0.36	0.0137
E	0	0.09	0.48	0.06	0.13	0.23	0.33	0.48	0.19	0.0725

Graph 2.6.1: Normalized relationship of progress groupings by school



Graph 2.6.2: Normalized progress grouping by School with Average represented



- Discretization (change of scale) : Applying categories alerting by progress by school (count of students)
    - Schools at risk of students not completing their lessons:
        - Groups: GREEN, YELLOW, RED
            - GREEN: Completed, Very Ahead, Middling
            - YELLOW: Behind
            - RED: More Behind, Very Behind

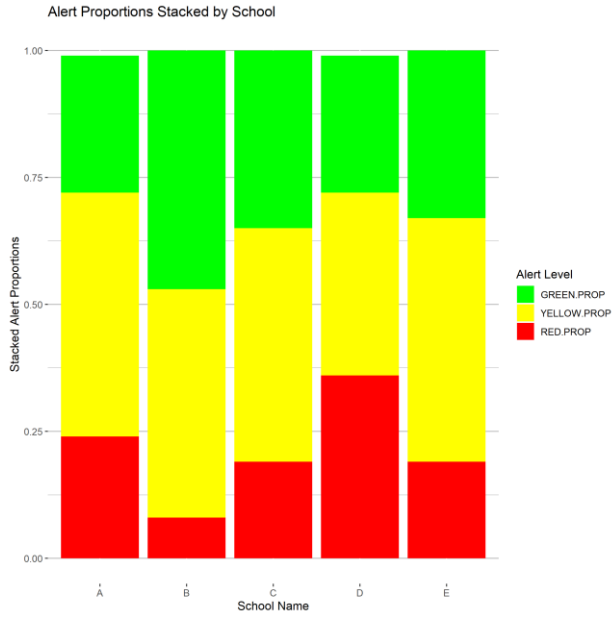
A new scale grouping was added to the proportion's dataset to further narrow down the feature set and group schools into categories that would indicate if they had a large population of students at risk of not completing all the lessons for the term.

Based on proportion of students, Graph 2.8.1 below visualizes how school 'D' has the highest proportion of its students at risk of not completing their lessons, while school 'B' has the lowest proportion. Whereas in contrast, Graph 2.8.2 blow, which visualizes alerts by total count of Students in the grouping per School, shows that School 'A' has the highest number of Students at risk of not completing their math lessons this term with School 'D' being the least.

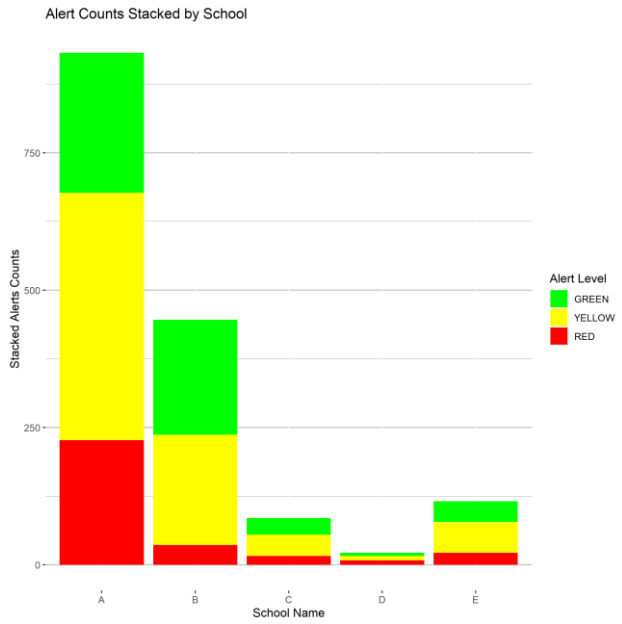
Table 2.8: Table view of school dataset as proportions to student count by school

School	VeryAhead.PROP	Middling.PROP	Behind.PROP	MoreBehind.PROP	VeryBehind.PROP	Completed.PROP	GREEN.PROP	YELLOW.PROP	RED.PROP	Total.PROP
A	0	0.12	0.48	0.08	0.17	0.15	0.27	0.48	0.24	0.5821
B	0	0.19	0.45	0.03	0.05	0.28	0.47	0.45	0.08	0.2786
C	0	0.13	0.46	0.05	0.14	0.22	0.35	0.46	0.19	0.0531
D	0	0.14	0.36	0.09	0.27	0.14	0.27	0.36	0.36	0.0137
E	0	0.09	0.48	0.06	0.13	0.23	0.33	0.48	0.19	0.0725

Graph 2.8.1: Progress Alerts as a Proportion by School



Graph 2.8.2: Progess Alerts by Total Count by School

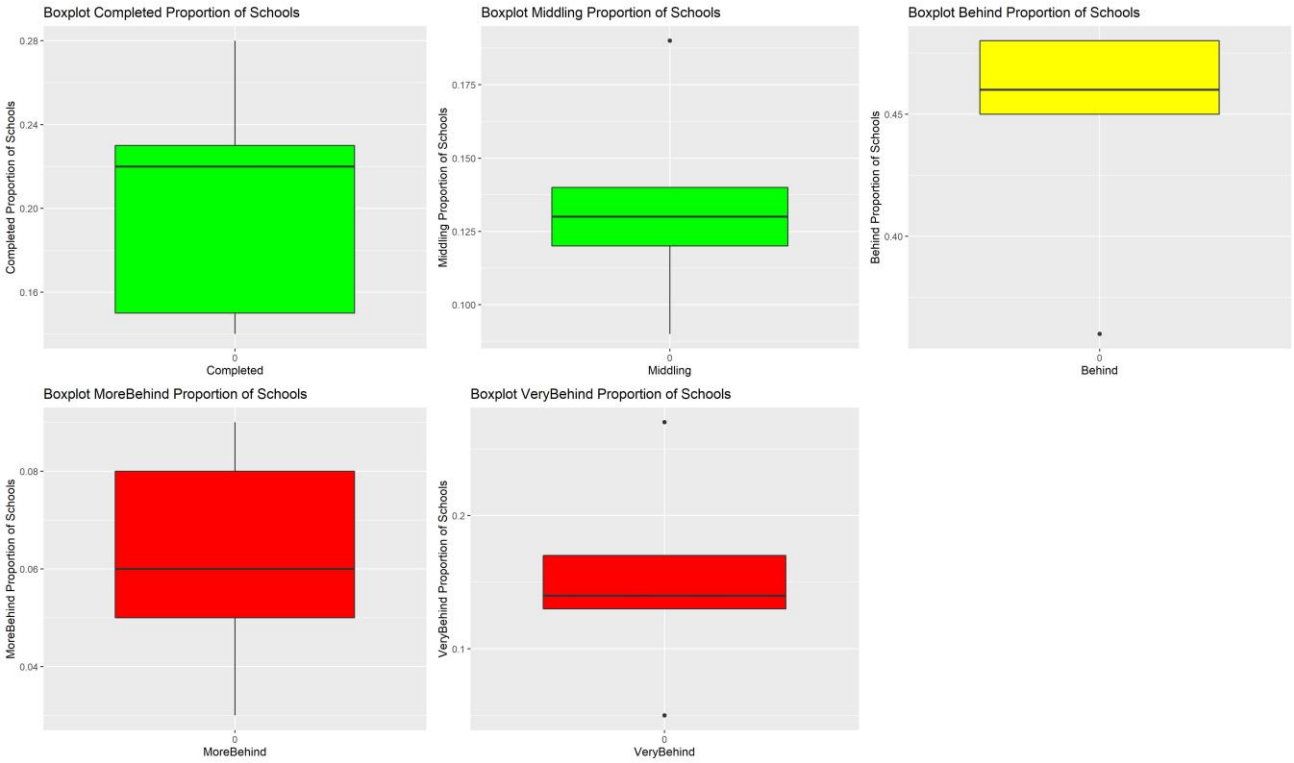


Homework Assignment 1 (week 1)

Table 2.9: Summary Table of School dataset as proportions

School	VeryAhead.PROP	Middling.PROP	Behind.PROP	MoreBehind.PROP	VeryBehind.PROP	Completed.PROP	GREEN.PROP	YELLOW.PROP	RED.PROP	Total.PROP
A:1	Min. :0	Min. :0.090	Min. :0.360	Min. :0.030	Min. :0.050	Min. :0.140	Min. :0.270	Min. :0.360	Min. :0.080	Min. :0.0137
B:1	1st Qu.:0	1st Qu.:0.120	1st Qu.:0.450	1st Qu.:0.050	1st Qu.:0.130	1st Qu.:0.150	1st Qu.:0.270	1st Qu.:0.450	1st Qu.:0.190	1st Qu.:0.0531
C:1	Median :0	Median :0.130	Median :0.460	Median :0.060	Median :0.140	Median :0.220	Median :0.330	Median :0.460	Median :0.190	Median :0.0725
D:1	Mean :0	Mean :0.134	Mean :0.446	Mean :0.062	Mean :0.152	Mean :0.204	Mean :0.338	Mean :0.446	Mean :0.212	Mean :0.2000
E:1	3rd Qu.:0	3rd Qu.:0.140	3rd Qu.:0.480	3rd Qu.:0.080	3rd Qu.:0.170	3rd Qu.:0.230	3rd Qu.:0.350	3rd Qu.:0.480	3rd Qu.:0.240	3rd Qu.:0.2786
NA	Max. :0	Max. :0.190	Max. :0.480	Max. :0.090	Max. :0.270	Max. :0.280	Max. :0.470	Max. :0.480	Max. :0.360	Max. :0.5821

Graph 2.9.1: Boxplot representation of each progress category



3 Conclusions

Based on proportion of students, shown in Graph 2.8.1 that visualizes how school 'D' has the highest proportion of its students at risk of not completing their lessons, while school 'B' has the lowest proportion; The data can be used to rank these five school's performances in relation to each other as:

- 1st Rank: School B
- 2nd Rank: School C
- 3rd Rank: School E
- 4th Rank: School A
- 5th Rank: School D

Keeping in mind however, this ranking is strictly based on proportion of students to the given school and does not represent the number of students who are impacted by these performance gaps directly. Assessing for overall impacts to student success rate of not completing the math lessons for the term would focus on school 'A' over all others. School 'A' has 60% of the overall student population, having just under 250 students at risk of not completing the course.