IST565 Data Mining

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 9/8/2019**

HW2 Instruction

This is a real job interview question from a data analysis company, and I doubt there is a standard answer to this question. So feel free to explore your story by using the data exploration and transformation techniques appropriately.

----------instruction quote begins-------------

Here is a small dataset for you to work with.

Each of 5 schools (A, B, C, D and E) is implementing the same math course this semester, with 35 lessons. There are 30 sections total. The semester is about 3/4 of the way through.

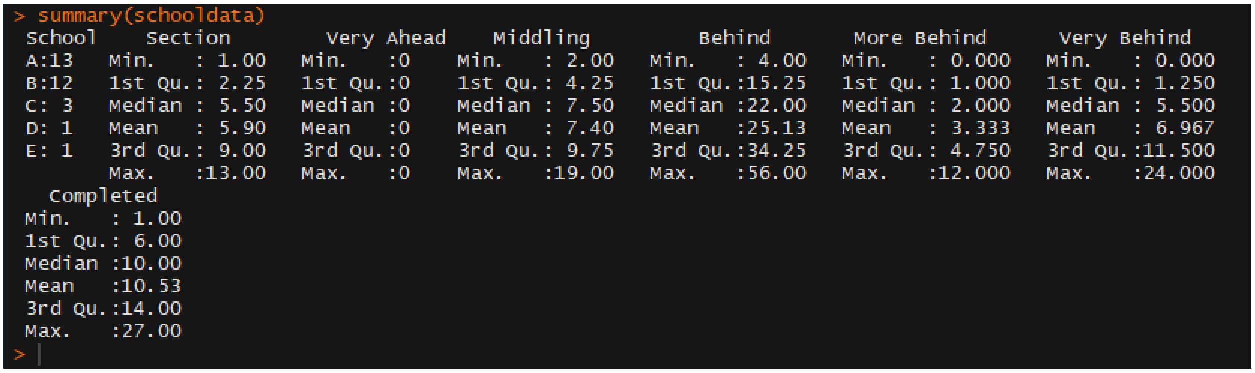
For each section, we record the number of students who are:

- 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)

What's the story (or stories) in this data? Find it, and tell it visually and, above all, truthfully.

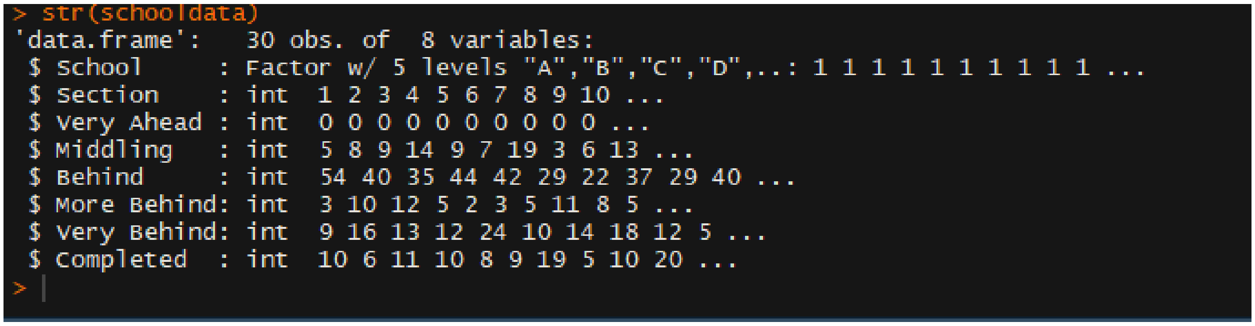
-----------instruction quote ends-----------------

```
> summary(schooldata)
```



```
> summary(schooldata)
 School     Section       Very Ahead     Middling          Behind        More Behind       Very Behind    
 A:13   Min.   : 1.00   Min.   :0    Min.   : 2.00   Min.   : 4.00   Min.   : 0.000   Min.   : 0.000  
 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  
 C: 3   Median : 5.50   Median :0    Median : 7.50   Median :22.00   Median : 2.000   Median : 5.500  
 D: 1   Mean   : 5.90   Mean   :0    Mean   : 7.40   Mean   :25.13   Mean   : 3.333   Mean   : 6.967  
 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  
        Max.   :13.00   Max.   :0    Max.   :19.00   Max.   :56.00   Max.   :12.000   Max.   :24.000  
   Completed    
 Min.   : 1.00  
 1st Qu.: 6.00  
 Median :10.00  
 Mean   :10.53  
 3rd Qu.:14.00  
 Max.   :27.00  
> 
```

```
> str(schooldata)
```

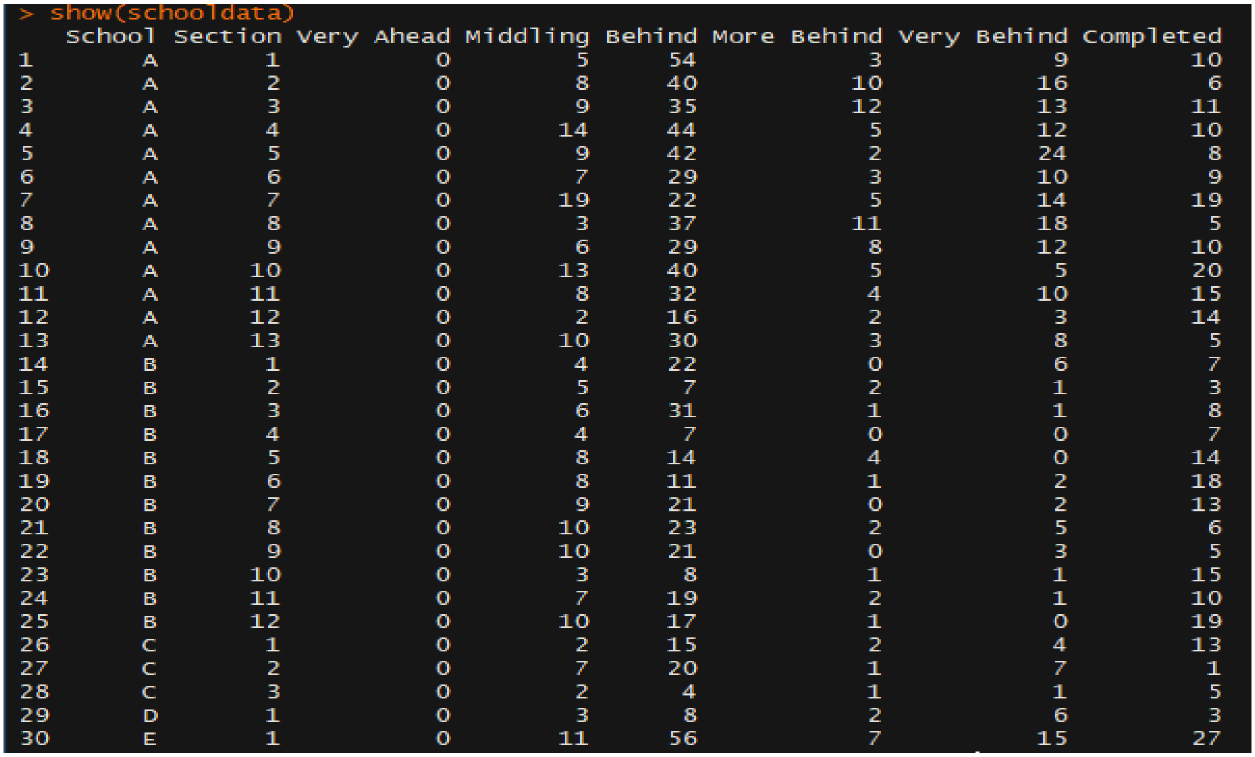


```
> str(schooldata)
'data.frame':	30 obs. of  8 variables:
 $ School     : Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Section    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Very Ahead : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Middling   : int  5 8 9 14 9 7 19 3 6 13 ...
 $ Behind     : int  54 40 35 44 42 29 22 37 29 40 ...
 $ More Behind: int  3 10 12 5 2 3 5 11 8 5 ...
 $ Very Behind: int  9 16 13 12 24 10 14 18 12 5 ...
 $ Completed  : int  10 6 11 10 8 9 19 5 10 20 ...
> 
```

Using **summary()** and **str()**, we can find out the summary and the structure of the given dataset

Here, we can see that all the variables are numeric, except for school.

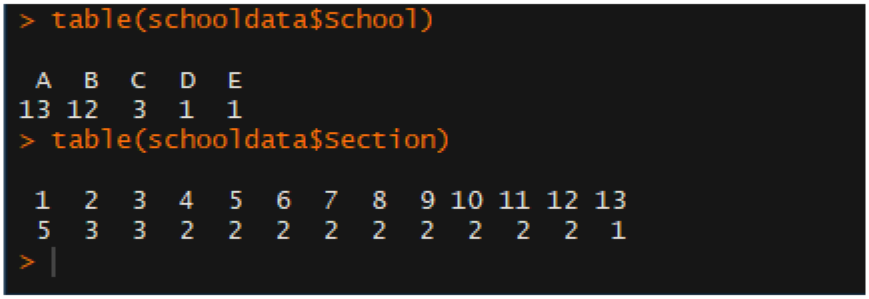
```
> show(schooldata)
```



```
> show(schooldata)
   School Section Very Ahead Middling Behind More Behind Very Behind Completed
1       A       1          0        5     54           3           9        10
2       A       2          0        8     40          10          16         6
3       A       3          0        9     35          12          13        11
4       A       4          0       14     44           5          12        10
5       A       5          0        9     42           2          24         8
6       A       6          0        7     29           3          10         9
7       A       7          0       19     22           5          14        19
8       A       8          0        3     37          11          18         5
9       A       9          0        6     29           8          12        10
10      A      10          0       13     40           5           5        20
11      A      11          0        8     32           4          10        15
12      A      12          0        2     16           2           3        14
13      A      13          0       10     30           3           8         5
14      B       1          0        4     22           0           6         7
15      B       2          0        5      7           2           1         3
16      B       3          0        6     31           1           1         8
17      B       4          0        4      7           0           0         7
18      B       5          0        8     14           4           0        14
19      B       6          0        8     11           1           2        18
20      B       7          0        9     21           0           2        13
21      B       8          0       10     23           2           5         6
22      B       9          0       10     21           0           3         5
23      B      10          0        3      8           1           1        15
24      B      11          0        7     19           2           1        10
25      B      12          0       10     17           1           0        19
26      C       1          0        2     15           2           4        13
27      C       2          0        7     20           1           7         1
28      C       3          0        2      4           1           1         5
29      D       1          0        3      8           2           6         3
30      E       1          0       11     56           7          15        27
```

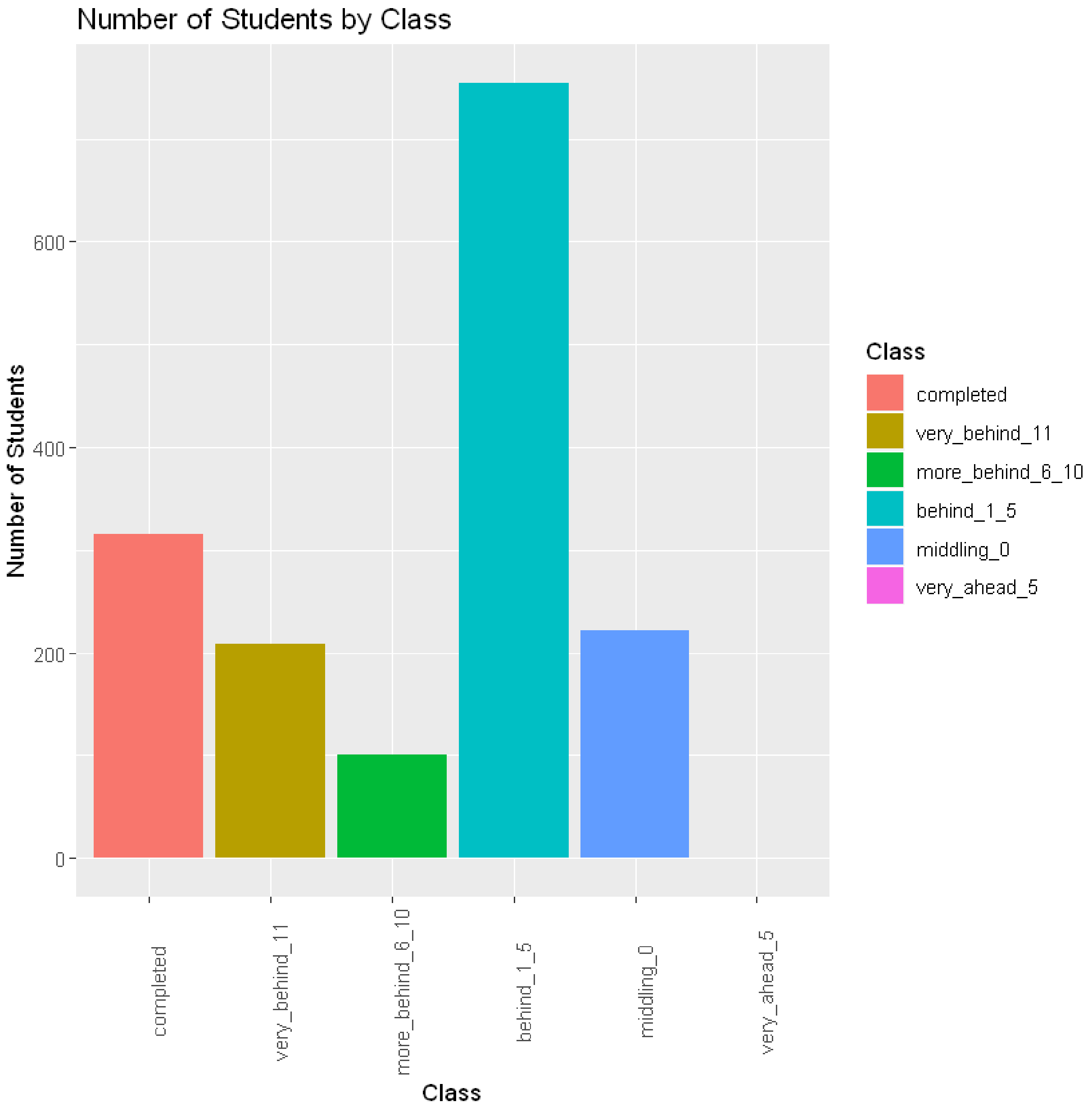
```
> table(schooldata$School)
```

```
> table(schooldata$Section)
```

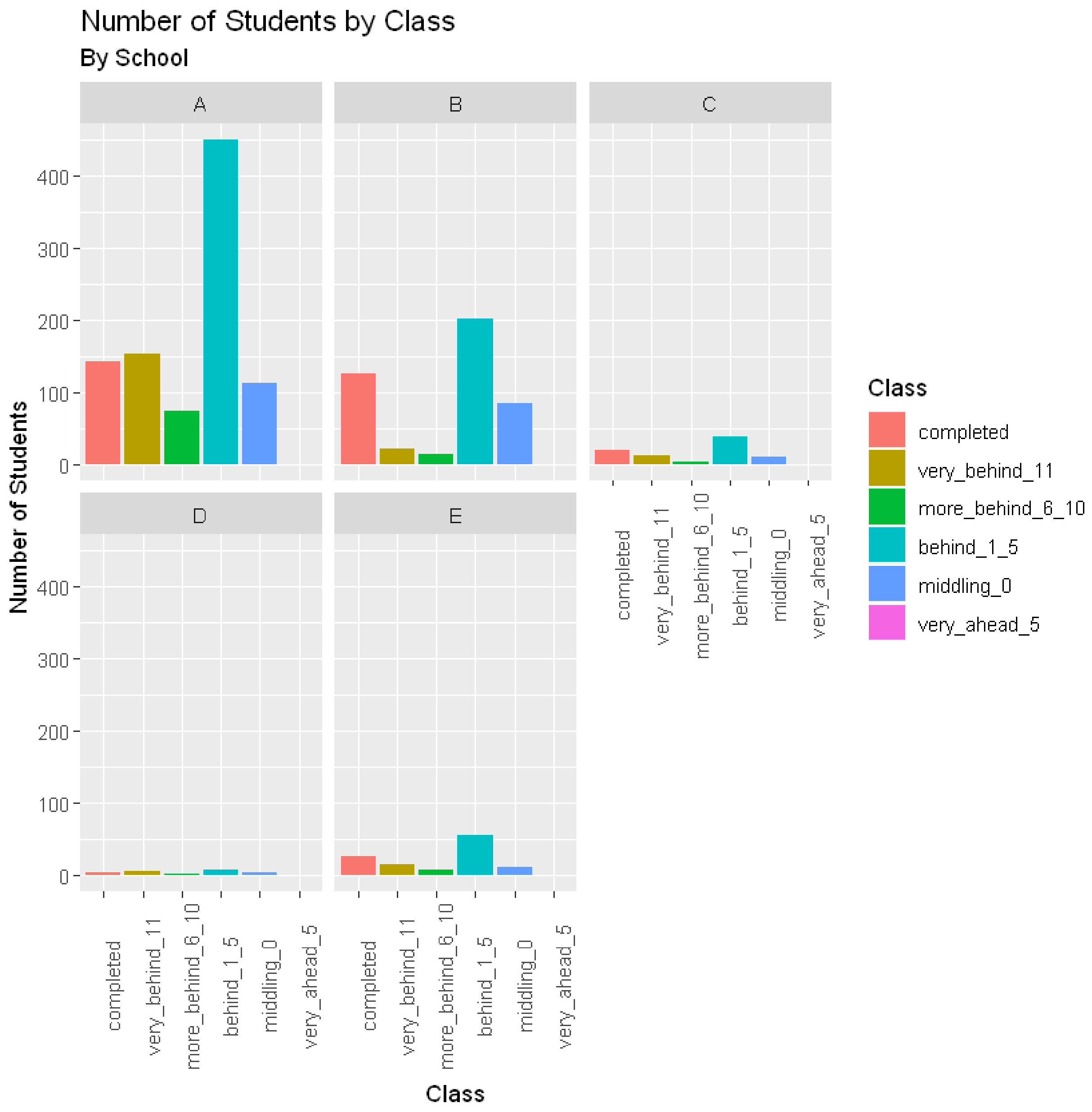


```
> table(schooldata$School)

 A  B  C  D  E 
13 12  3  1  1 
> table(schooldata$Section)

 1  2  3  4  5  6  7  8  9 10 11 12 13 
 5  3  3  2  2  2  2  2  2  2  2  2  1 
> 
```

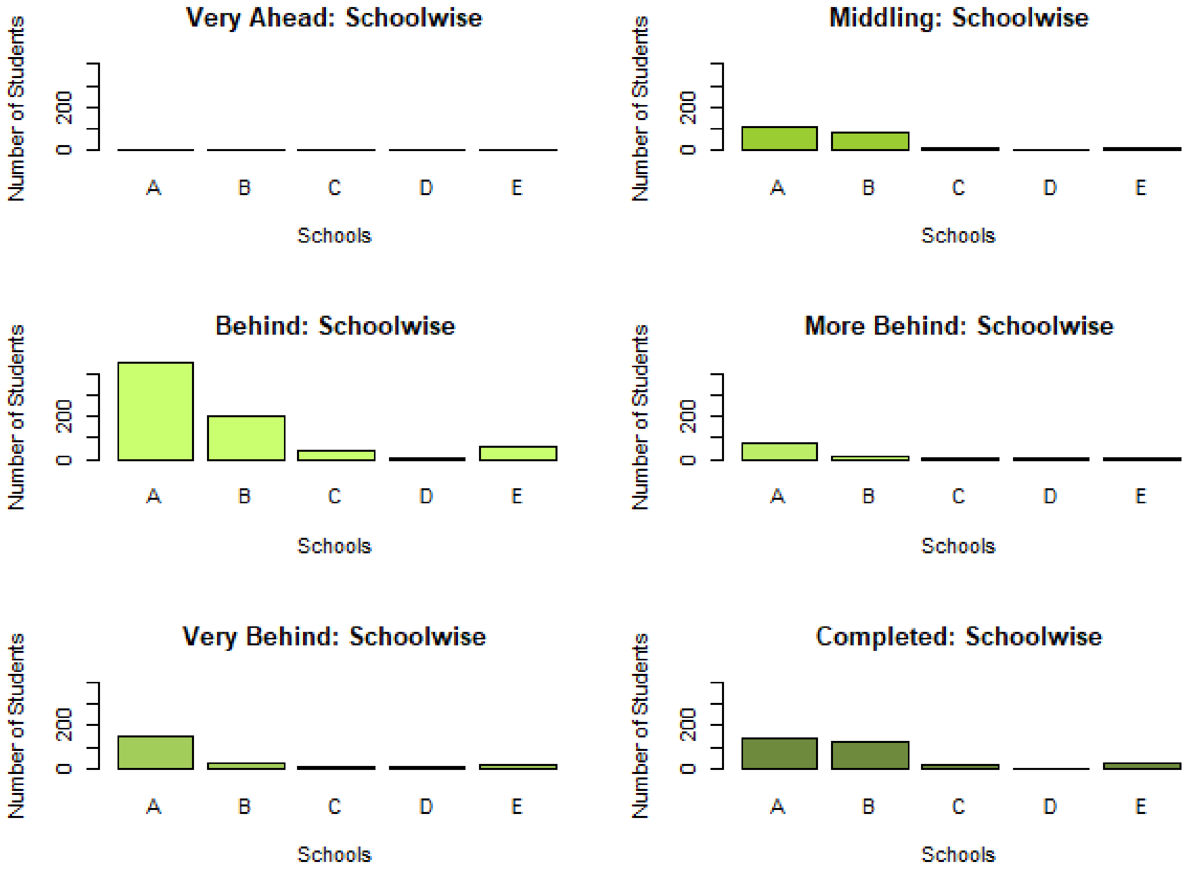
Here we can see the number of math section across each school. School A has 13 sections, school B has 12 sections, school C has 3 sections, school D and E have 1 section each,



Here we can see that more students (approximately 750) are lagging by 1-5 lessons whereas number of students (approximately 100) lagging by 6-10 lessons is comparatively less. We can also see that no student is very ahead (by5 lessons).



Here, we can see the same thing as earlier. More number of students are in range of behind by 1-5 lessons.



No school has students who are “Very ahead”

A is ahead of other schools in almost every other category, probably because **it has comparatively more students**.

Number of students in school C, E and D are very less.

Conclusions:

1. The number of students across schools are:
   A > B > C > E > D
2. The number of students who are behind (1-5 lessons behind) are substantially higher than the other categories.
3. There are no students who are very ahead in the course, in any school.
4. Most of the students are 1-5 lessons behind, across all the schools.
5. Following is the order of sizes of levels of completion:
6. Behind > Completed > Middling > Very Behind > More Behind > Very Ahead