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)

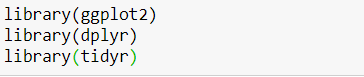
• completed (finished with the course)

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

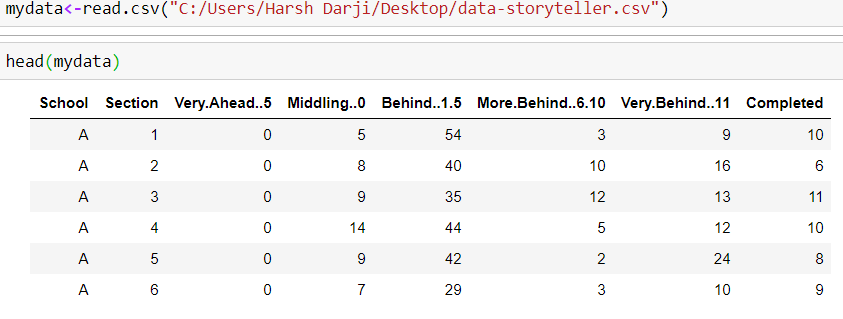
-----------instruction quote ends-----------------

**Answer**

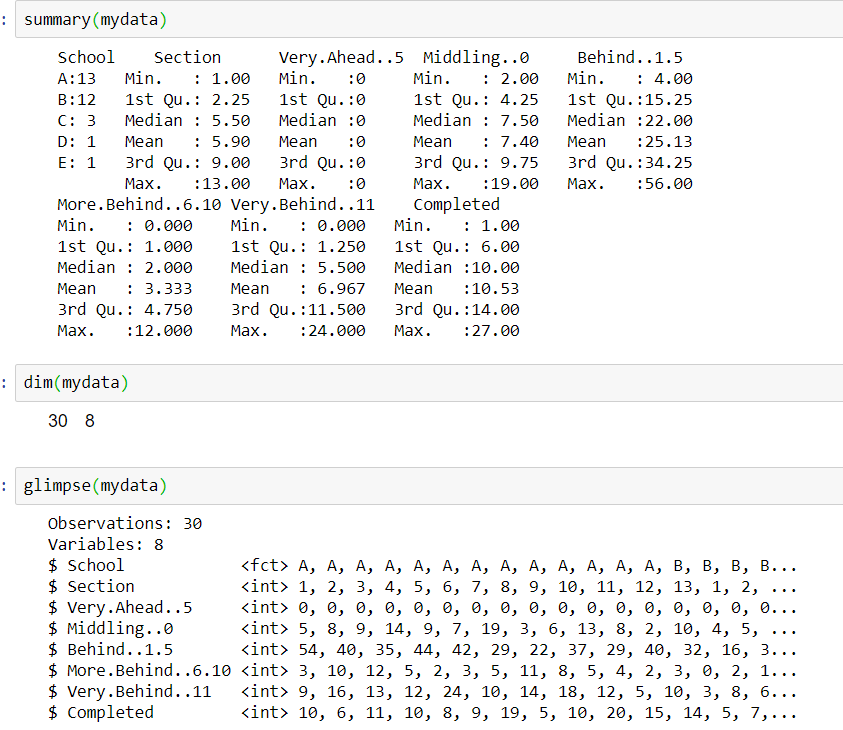
Loading the required packages in R to restructure and Visualize the given data.



Reading the dataset and displaying it in a tabular form.



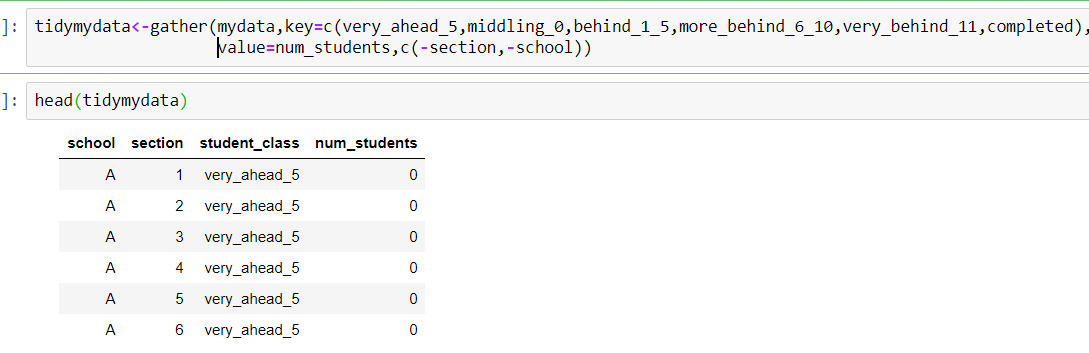
Now, let’s find out the structure of data and the amount of data we have to understand it better.



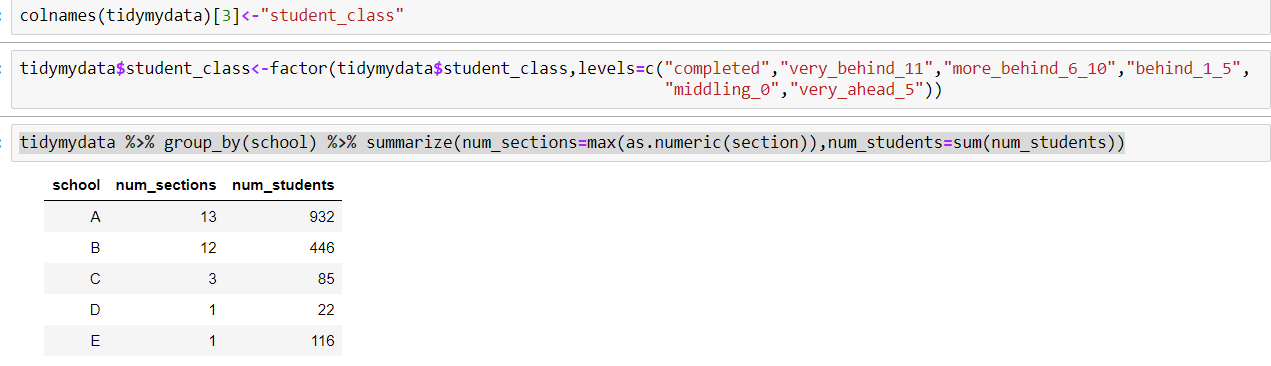
We see that we have 30 rows and 8 columns. We rename the columns so that we simplify them and it will be easy to us them.



We now re-structure the data such that we have one row per observation using tidyr package.

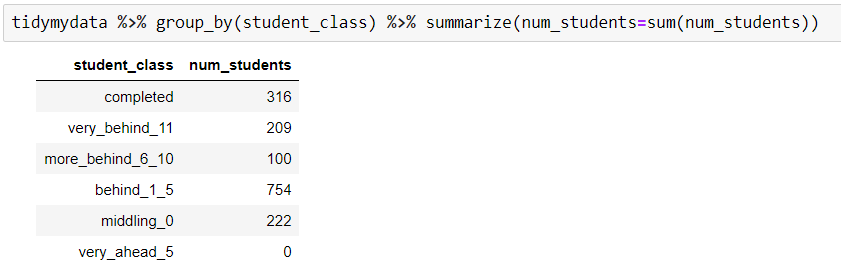


Grouping the data base don the schools.

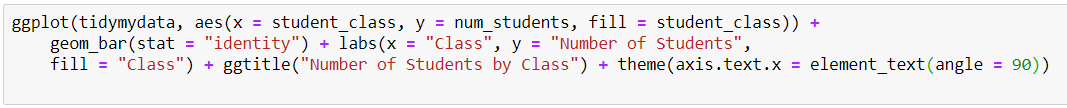


Here, we observe that though school A and school B have similar number of sections, number of students in school A is almost double than school B. Also, school D has only one section and least number of students.

Now, let’s check out the data by the number of lessons they have completed i.e by their class.



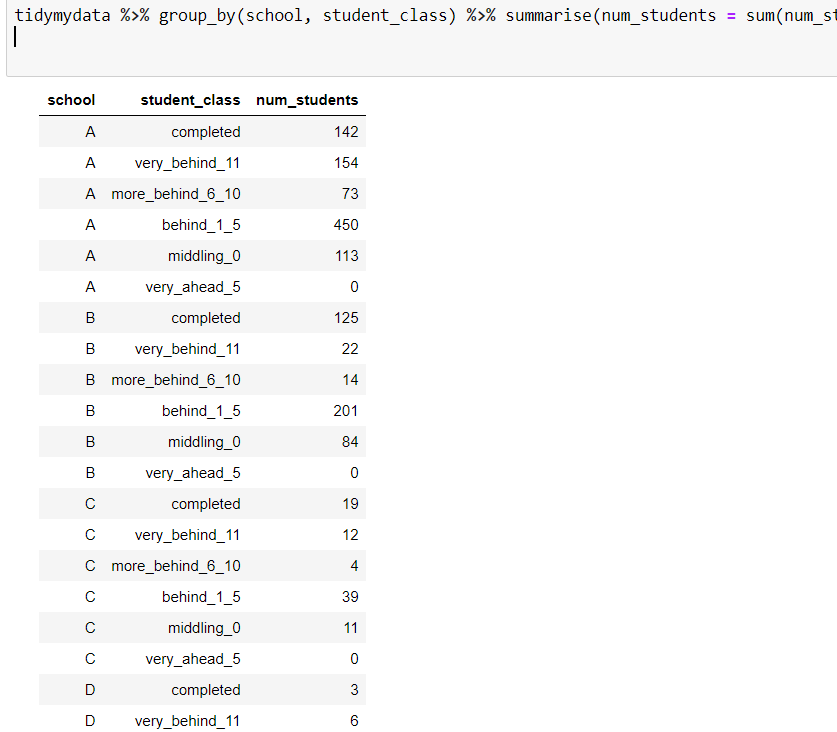
Here, we observe that Most of the students are behind by 1-5 lessons and no one is 5 lessons ahead. Let’s visualize to understand the scenario better.



A screenshot of a cell phone

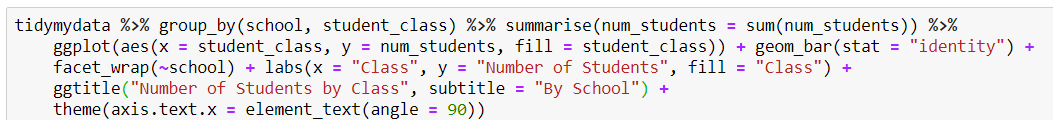
Description generated with high confidence

Here, the conclusion is same, more students are lagging behind by 1-5 lessons. We break down our data to get more insights. We group our data by school and class of the student.



We observe that 450 students are behind by 1-5 lessons which is a lot. School A has 72.6% behind rate while other schools have 57.7% behind rate. Also, school A has 15.2% completed rate while other schools have 26% complete rate. Is there a problem with School A? or Is complete rate dependent on number of sections and Number of students? Let’s find it out.

Now, let’s check the scenario of students and their class for each school.

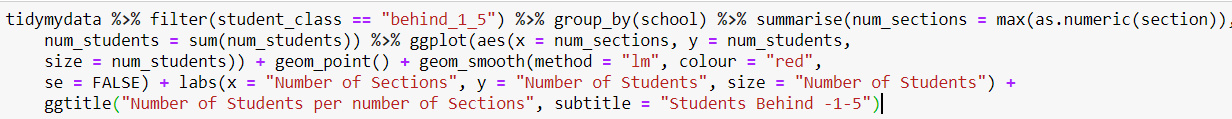


A screenshot of a cell phone

Description generated with very high confidence

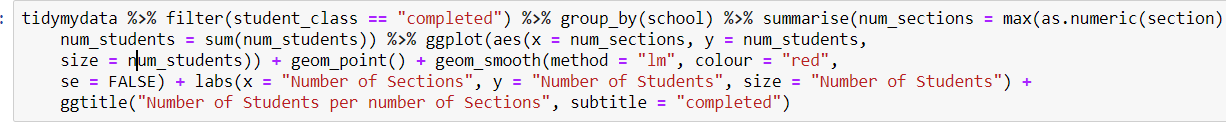
We observe that in most of the schools, number of students are behind by 1-5 lessons are highest. While in school D number of students behind by 1-5 lessons ad very behind(11 lessons\_ are similar)

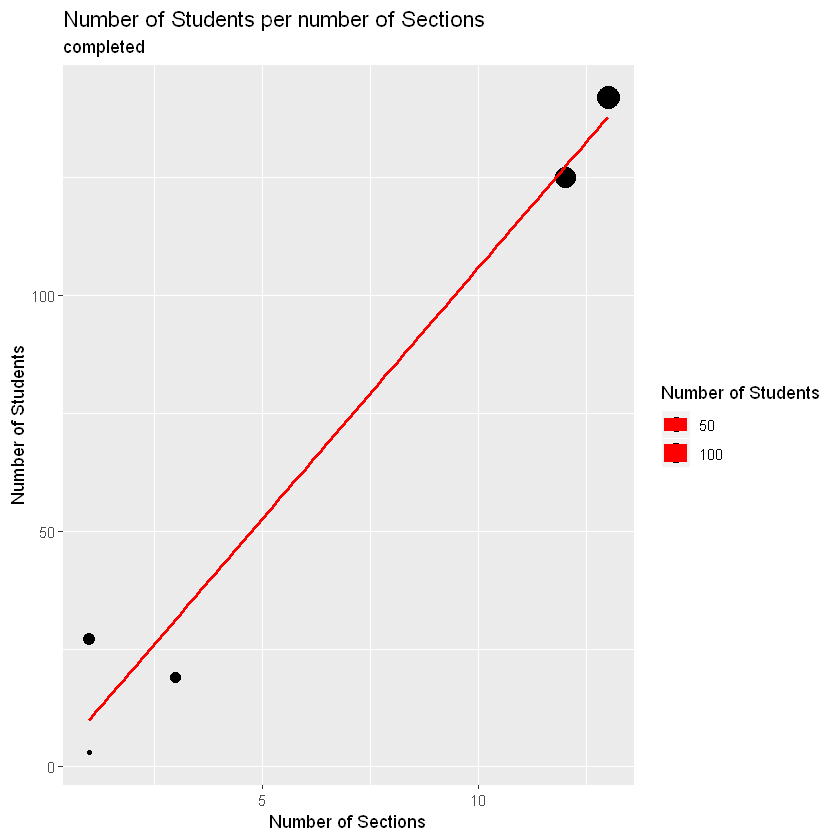
Let’s find out the relation between number of sections and number of students for each class:

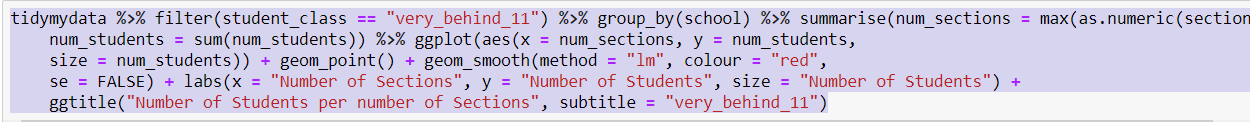


A close up of a map

Description generated with high confidence

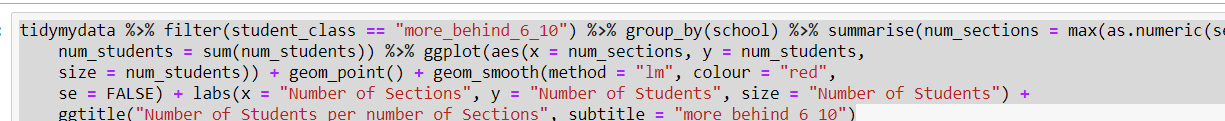






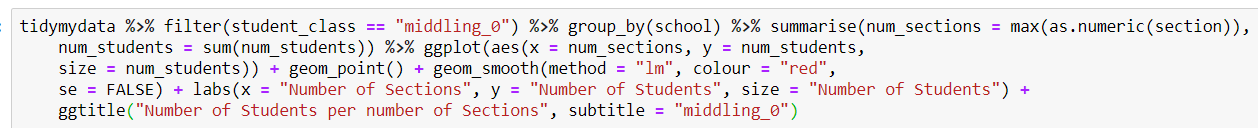
A screenshot of a cell phone

Description generated with high confidence



A screenshot of a cell phone

Description generated with high confidence

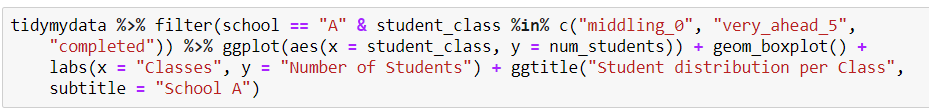


A screenshot of a cell phone

Description generated with high confidence

It’s evident that there’s a relation between number of sections and number of students lagging behind Math lessons as well as number of students that have completed. But data set is quite small to conclude that these two entities are correlated.

Let’s look at the students of school A who are behind to understand the relation.



A close up of a logo

Description generated with very high confidence

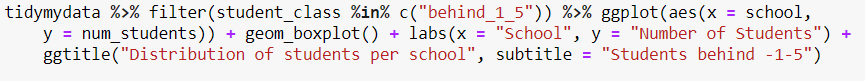
there seem to be no outliers that are driving School A to have such high numbers of students lagging behind. However, we can say that the largest section has nearly 55 students lagging behind while the smallest section has a bit over 15. With a median of 35 students, this explains that the number of sections is driving school A's high numbers.

Now, let’s look at the scenario of completed students.

A close up of a logo

Description generated with very high confidence

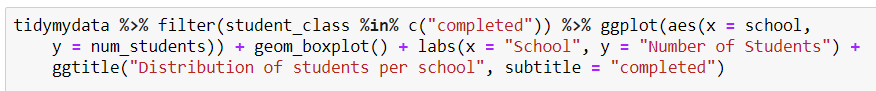
Student that have completed are very less 20 for larger section while smaller section have only 5 students that have completed Math lesson.



A screenshot of a cell phone

Description generated with very high confidence

So, there’s no real explanation why school A is driving up the numbers other than they have more students and sections - something evident by the fact that school E has one section with 56 students lagging behind by 1 to 5 courses. Other than that, there are no outliers nor anything that jumps out when trying to explain why class A has more students lagging above the average



**A screenshot of a cell phone

Description generated with very high confidence**

School E has most number of students that have completed Math lessons.