Matthew L. Pergolski
IST 707
1/27/2022

Introduction

The following reviews the state of a school district that contains five (5) schools in a particular semester. For the purposes of privacy, no school names will be directly shared. Instead, the intent of report is to provide analysis of how students are performing (anonymously) across schools within a district. Upon the completion of the report, a reader/viewer will be able to review a performance ranking of each school and know which school is performing the best/worst among peers.

Audiences interested in the findings of the report may include (but are not limited to) any of the following:

- Students
- Parents
- Teachers
- Local and national government officials/representatives
- Colleges and universities
- Sponsors

Consequently, information gained from the report may also serve as a reference for schools to provide additional resources for their students in hopes to drive an increase in academic performance.

Analysis and Models

About the Data

As previously mentioned, the names of the schools within the district are not specifically mentioned in the public report. Alternatively, the stakeholders mentioned above will be notified of subsequent instructions to provide transparency on where the children's performance stands.

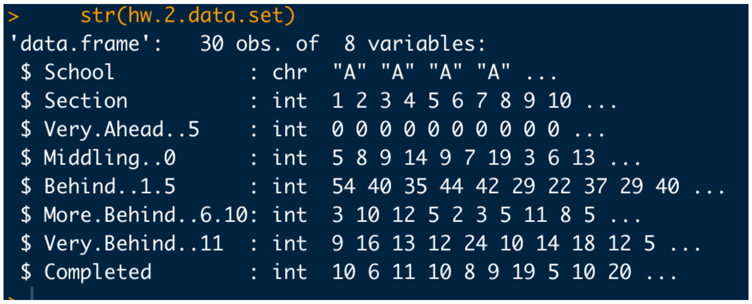
Each of the five (5) schools are labeled A, B, C, D, and E, respectively. The data shows performance relative to a math course during the term, which is already approximately three quarters of the way complete. 35 lessons span across the class, with 30 sections total.

Across the many sections, the performance of the students are 'bucketed' into 6 different categories. They appear as follows:

- 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

Upon loading the data from the downloaded CSV file from 2SU, a data frame is shown with 30 observations and 8 variables.



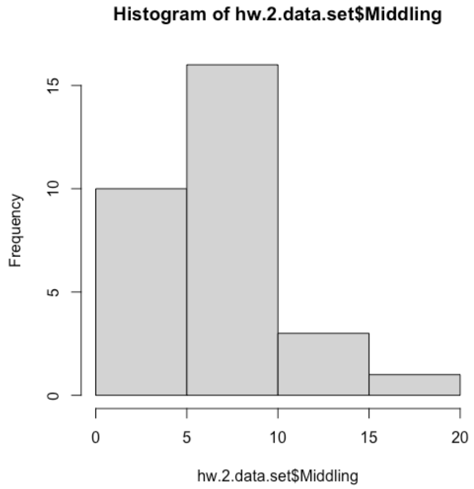
```
>     str(hw.2.data.set)
'data.frame':	30 obs. of  8 variables:
 $ School           : chr  "A" "A" "A" "A" ...
 $ Section          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Very.Ahead..5    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Middling..0      : int  5 8 9 14 9 7 19 3 6 13 ...
 $ Behind..1.5      : int  54 40 35 44 42 29 22 37 29 40 ...
 $ More.Behind..6.10: int  3 10 12 5 2 3 5 11 8 5 ...
 $ Very.Behind..11  : int  9 16 13 12 24 10 14 18 12 5 ...
 $ Completed        : int  10 6 11 10 8 9 19 5 10 20 ...
```

Among the eight (8) variables shown, six (6) appear to be very familiar, as they are the ‘bucketed’ categories mentioned earlier. Each ‘bucket’ associates with a performance indicator to classify where students fall in overall academic performance. Each were imported as a numeric type, so data ‘munging’ will not be required for these attributes.

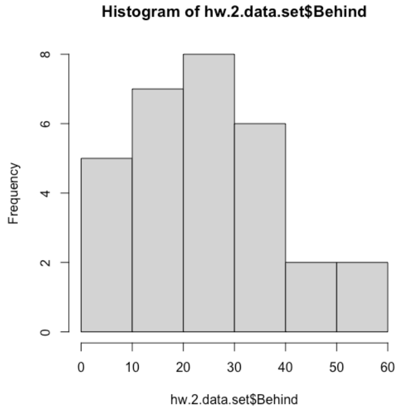
The first two (2) variables of the data frame relate to the school alias name (i.e., A, B, C, D, or E) and the sections for each respective school. The ‘School’ variable is currently listed as a character, which will be turned into a factor by the determination of the author – the same will be done for the ‘Section’ variable. Both attributes can be seen as ‘categorical,’ and thus would be best suited to be tied as an ‘ordinal’ factor.

In scoping through the rest of the data via the summary() and View() commands, no missing values were detected; the variable headers appear to have formatting issues, so R code was developed to ‘clean’ the attribute names in order to provide a more professional format to the reader. Further, it does appear that as a specific variable – ‘Very Ahead’ – has 0 for all observations. It will be kept in the data set but may not be useful in prospective calculations.

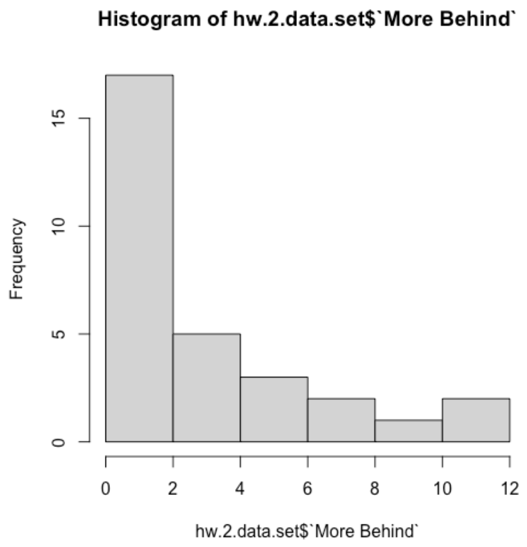
When looking at each numerical variable through a histogram, the following is seen (Note: ‘Very Ahead’ variable will not be shown due to ‘zero’ data):



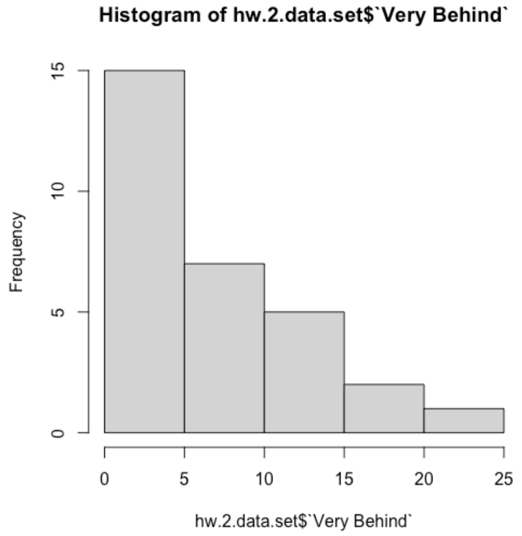
For the 'Middling' attribute, a positively skewed distribution is seen – with a slight tail extending to the right of the graph. A median value for this variable equates to 7.5. The median was chosen over the mean, since the mean is more 'sensitive' to potential outliers.



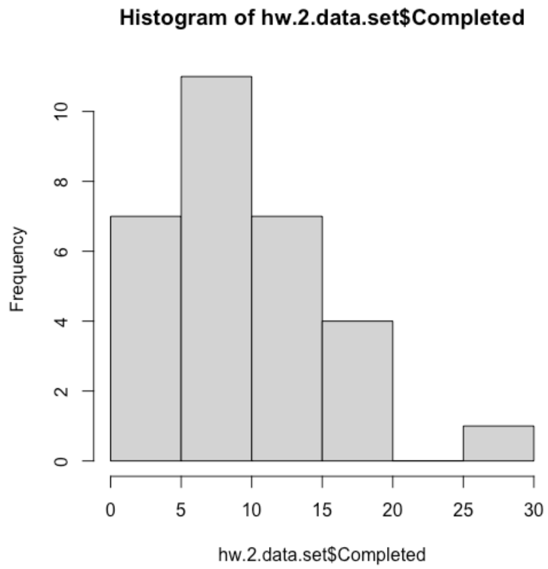
For the 'Behind' attribute, something closer to a normal distribution is witnessed, compared to the 'Middling' variable. The median value equates to 22.



The ‘More Behind’ variable conveys a definite positively skewed distribution, with a long tail extending to the right of the graph. The median value of is 2.



The ‘Very Behind’ attribute also appears to have a positively skewed distrubtion, with a median value of 5.5.

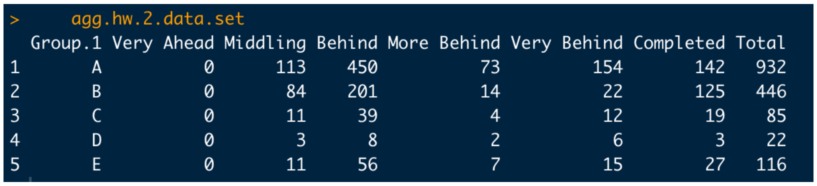


The ‘Completed’ variable seems to also be positively skewed – with a notable outlier. The median value is set at 10.53 (repeating).

Exploring the Data

Although these variables are interesting to look at – especially one by one – there are limitations that exist in the current format. First, the audience views the data as a whole. In other words, there is not currently visibility into differentiating each school by variable. Consequently, the aggregate() function will be done to group by school.

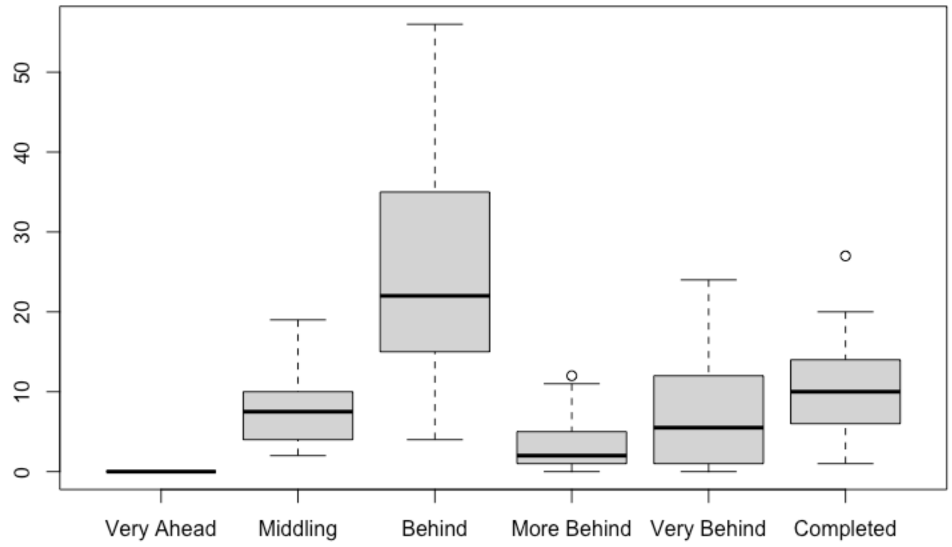
As a result, the newly condensed data set looks as follows:



```
>    agg.hw.2.data.set
  Group.1 Very Ahead Middling Behind More Behind Very Behind Completed Total
1       A          0      113    450          73         154       142   932
2       B          0       84    201          14          22       125   446
3       C          0       11     39           4          12        19    85
4       D          0        3      8           2           6         3    22
5       E          0       11     56           7          15        27   116
```

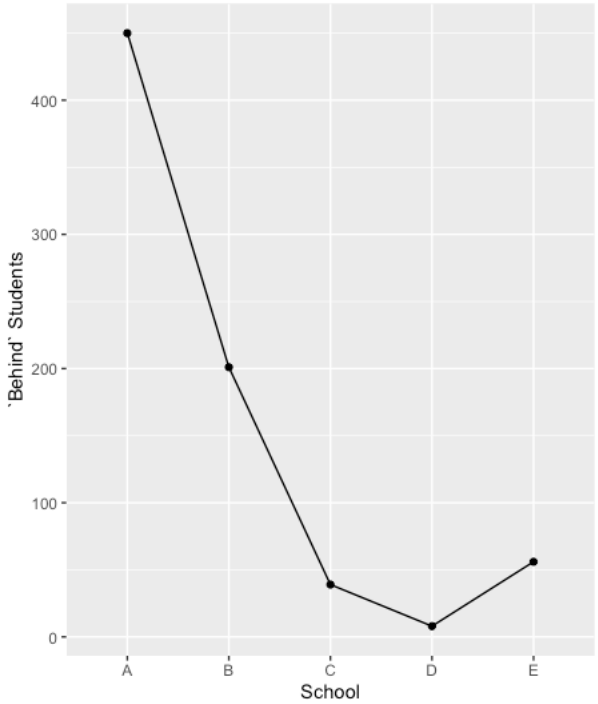
A ‘Total’ column was added to provide a total for each respective school.

When further exploring the original, uncondensed data in boxplot format, the following was observed:



From the uncondensed data, the ‘Behind’ variable seems to have the most dispersion compared to the alternatives. As a simplified ranking system used to determine which schools are ‘best’ through ‘worst,’ this will be determined by how many students are ‘Behind’ in their classes. This will shed light on which school’s have the most students behind in their classes – and will provide opportunity for future research to improve this outcome.

A basic line chart on the condensed, aggregated data set shows the number of students who are ‘Behind’ – this scatter plot conveys the relationship between the various schools:

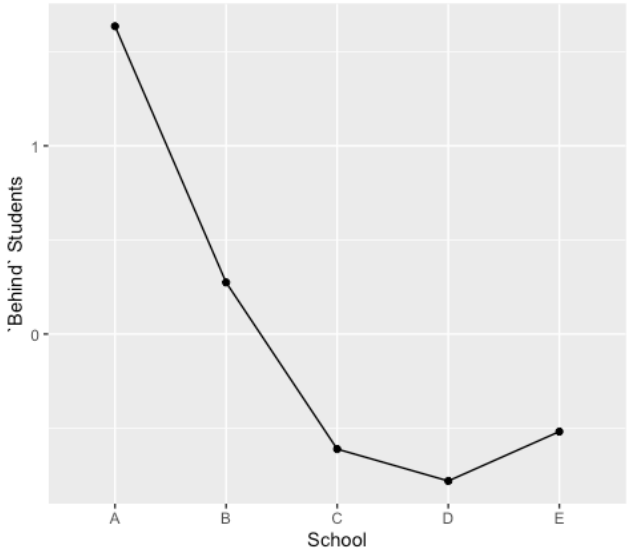


From the raw results, it would appear that school 'A' would be the clear worst institution, by the author's determined test, since it possesses the most students who are 'behind' in classwork; however, the results currently shown may be disproportionate due to the differences in total enrolled students. In other words, the data needs to be scaled in order to proportionate the data.

After executing the scaling of the aggregated data, the following can be observed in the results section:

Results

Although the data was scaled, similar results are seen. To scale the aggregated data, each value is subtracted by its respective attribute mean, then divided by its standard deviation. Subsequently, each variable's mean equates to zero while the standard deviation equals 1. By executing this scale command, all variables are proportionate to each other, leveling the playing field.



By interpreting the illustration, school 'A' remains the highest with a plotted value above 1, followed by school 'B, which is also positive. Schools 'E', 'C', and 'D' all had negative values after scaling.

Conclusions

As the scaled results illustrate, based on the determining factor of how many students are 'behind' at this point through the term/semester, the ranking for each school, from best to last, shows as follows:

1. D
    a. Least students 'behind' in coursework
2. C
3. E
4. B
5. A
    a. Most students 'behind' in coursework

To reiterate, the main factor for determining the ranking within was how many students were behind in their coursework. The 'Behind' attribute had the most dispersion as seen in the earlier boxplot of the uncondensed data.

From the findings in the report, although it appeared as though many students were 'behind' in their coursework across all schools within the district, it became known that school 'A' had the most significant role in this statistic, despite its larger student population.