Parin Patel

Data Exploration Mini-Project: Mathematics Curriculum

Introduction:

Pie Town, New Mexico has decided to initiate a slow rollout of its new high school mathematics curriculum. What started out as a simple idea proposed by a student running for Student Body President in one of the town's five high school, has turned into a widely petitioned initiative which majority of the town's residents support. A survey conducted by the school board on the matter showed that 80% of the town's residents believed that "mathematics was an important life skill". However, interestingly, only 40% of them felt "confident in their math ability". Additionally, majority of the towns recent high school graduates felt that "school did not adequately prepare them for life after graduation".

Therefore, this past fall, the school board voted to allow a standard, single-semester pilot course to take place in all five of the town's high schools. This new mathematics course will have 35-lessons and will incorporate technology to introduce students to common mathematics software's and programming languages. With a focus on developing core math skillsets, rather having a primary grade-based focus through homework and reading assignments, this new mathematics course's goal is to aid student development by including real-world applications.

With ¾ of the semester completed, the school board is trying to understand where the new mathematics classes stand in relation to the originally planned curriculum. The results of this data analysis will help them understand if the curriculum needs to be shortened or condensed even further. Additionally, it will also help them understand which schools in their district have students that are more open to participating in classes that follow a new mathematics curriculum.

Analysis:

The Data:

The dataset contained a spread from five high schools of how far a student had progressed through the new curriculum. In total, there are 30 sections between all the high schools in Pie Town. In each section, individual student progress of the new curriculum's material was recorded as one of the following:

Very Ahead (over 5 lessons ahead)
Middling (5 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)

Each of the sections reported the total number of students that were in each group. The dataset was put together after each section reported their student progress and included the section number and school.

Parin Patel

Data Preparation:

The working directory is set, and the original file is surveyed to confirm that it is in a compatible format for R (.csv). The data is then read and viewed to confirm the information matches the description provided. In this case, it is confirmed the data contains 30 sections across 5 schools with 8 total variables.

```
> head(storytellData, n=8) ##check data
  School Section Very.Ahead..5 Middling..0 Behind..1.5 More.Behind..6.10 Very.Behind..11 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
```

Figure 1.1

Next, the column names are formatted to remove unnecessary periods, spaces, and fillers.

```
> head(storytellData)
  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
```

Figure 1.2

Next, a check for missing values is conducted. The results of the check show that no missing values are found in the data.

```
> #check missing values
> msTotal<-sum(is.na(storytellData))
> cat("The number of missing values in the data is", msTotal)
The number of missing values in the data is 0
```

Figure 1.3

Next, the structure of the dataset is assessed to make sure each variable has the correct corresponding data type associated.

```
> str(storytellData)
'data.frame':    30 obs. of  8 variables:
 $ School     : Factor w/ 5 levels "A","B","C","D",..: 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 ...
```

Figure 1.4

It is found that the variable “Section” does not have the correct data type connected. Since the data type should be “factor,” a fix must be conducted. “Section” is the only variable with an incorrect data type. Therefore, the fix will only be done once.

```
> #change 'section' data type to factor
> storytellData1$Section<-as.factor((storytellData1$Section))
> str(storytellData1)
'data.frame':    30 obs. of  8 variables:
 $ School     : Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1...
 $ Section    : Factor w/ 13 levels "1","2","3","4",..: 1 2 3 4 5 6 ...
 $ 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 ...
```

Figure 1.5

The data will now be cleaned. This is done by looking through the variables and checking for a proper range in values. The goal of this check is to confirm the values in each of the variables refer to the number of students and that there are no negative numbers in the dataset. The lowest number possible in the dataset, after cleaning is 0. No negative values were found.

Data Processing:

Now that the data is cleaned, it is important to explore and process the information to understand what questions and patterns can be answered. At this stage, separate tables are formed from the original dataset. This will help to focus the scope of each analysis for a deeper understanding from the data.

First, the data is grouped by schools and the number of math-class sessions held at each school. The number of sessions at each school is aggregated to focus the information on counting the number of sessions. The table is then visually plotted as a bar graph. It results show that school A has the highest number of sessions, with 13 new math-classes held this semester. However, school B is not far behind with 12 new math sessions. School’s D and E both only have 1 new math course.

```
> storytellData_schoolAgg <-storytellData1 %>%
+   group_by(School) %>%
+   summarise(count=n())
> storytellData_schoolAgg
# A tibble: 5 x 2
  School count
  <fctr> <int>
1      A    13
2      B    12
3      C     3
4      D     1
5      E     1
```

Figure 2.1

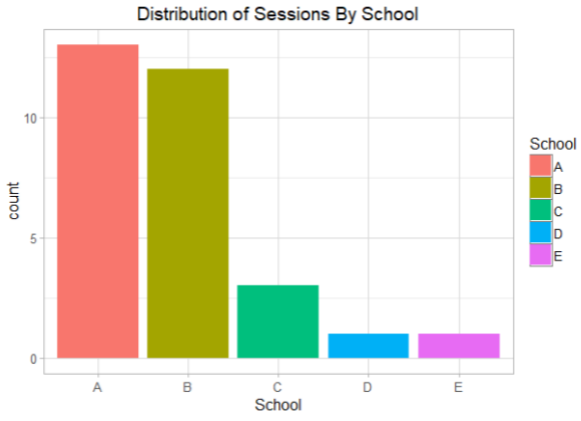

Figure 2.2

While the spread of schools shows a general overview of where the new math classes are held, the large disparity between the number of classes held at Schools A/B and Schools D/E stands out. Therefore, an analysis of the number of students taking the new pilot-math course is looked at.

```
> StudentEnrollSummary
  School NumStudents AverageStudent
1      A         932             72
2      B         446             37
3      C          85             28
4      D          22             22
5      E         116            116
```

Figure 3.1

The results of the student enrollment show that while, as expected, Schools A and B have the highest enrollment of students in the new pilot math class, School E has the highest number of average students per class. The distribution of students by school is shown on the plot below. The pattern of the students by school is like the pattern for the distribution of the number of sessions at each school.

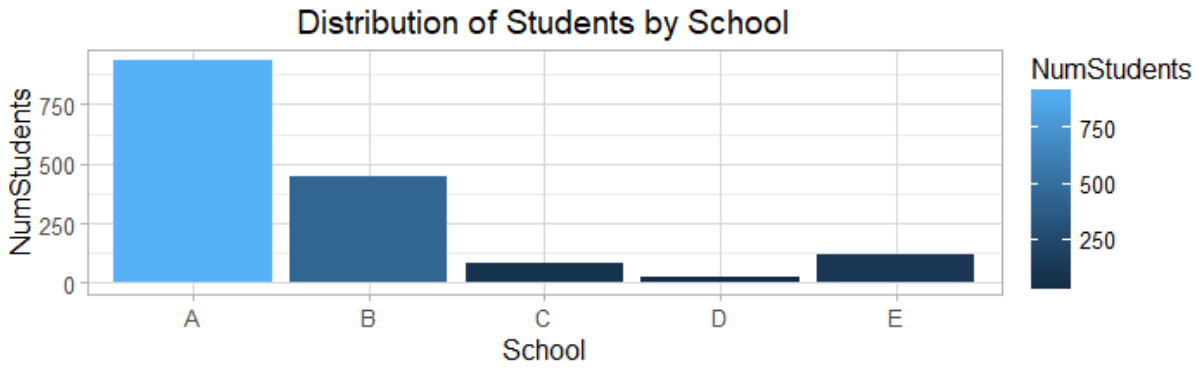

Figure 4.1

After a summary look at the schools and sessions are conducted, a deeper look at the variables can occur. The boxplot below shows the general spread of student progress for all schools. This method is used to compare the variables and detect any outliers.

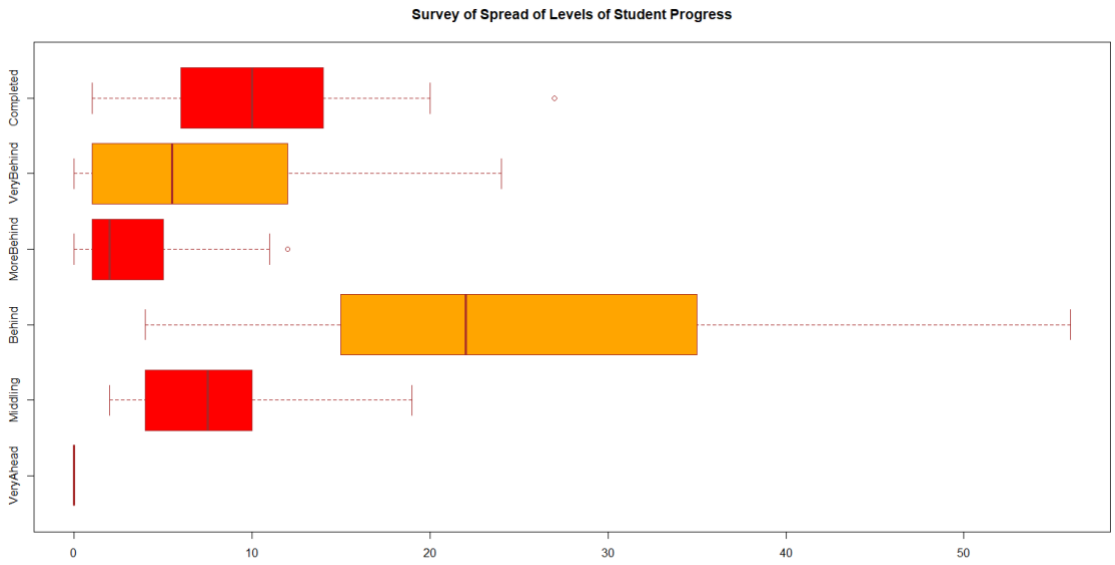

Figure 5.1

The boxplot above shows that more students reported being behind in the new pilot-math class. This value is higher than any other category. No students were reported being Very Ahead of the course, while an average of about 10 students reported finishing the course. It is also shown that more students reported being Very Behind than Completed in the new class. The boxplot also shows outliers in Completed and More Behind.

To analyze the learning status of the new curriculum at each school, scatterplots were created. The Schools initially split into separate tables. The tables were cleaned and then melted to prepare for the plot.

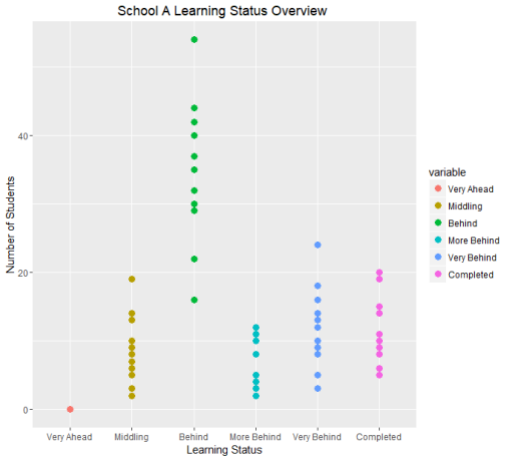

Figure 5.1

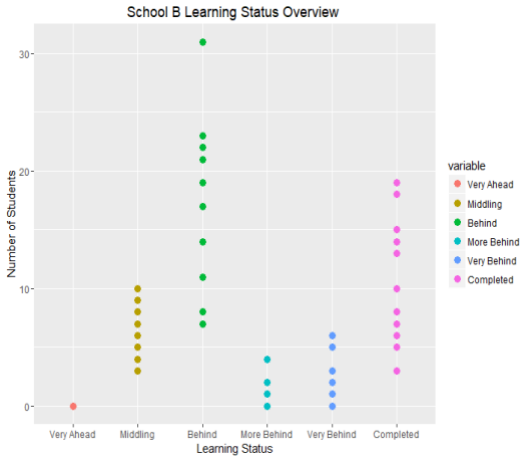

Figure 6.1

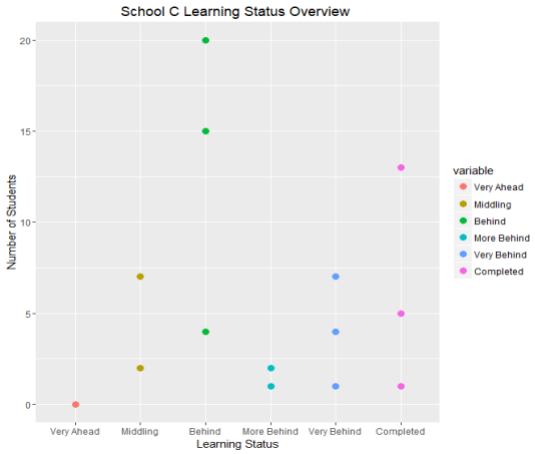

Figure 7.1

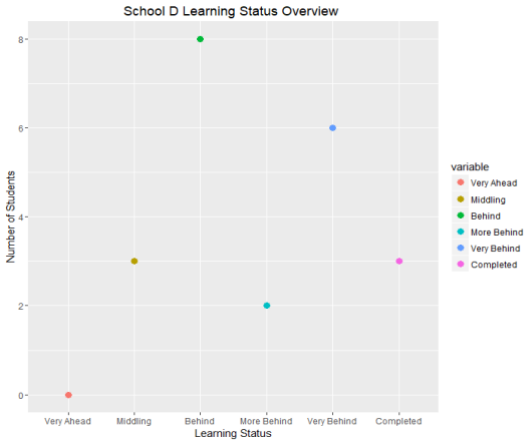

Figure 8.1

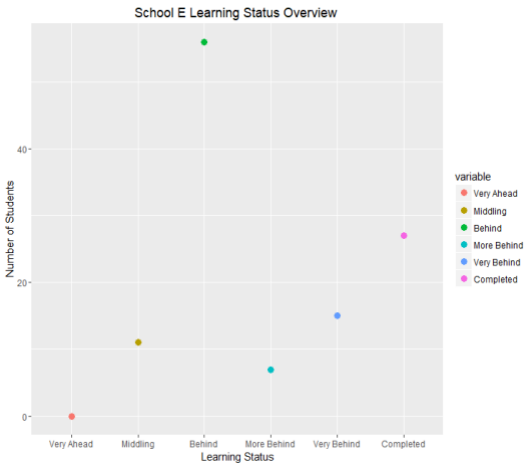

Figure 9.1

Results:

The results of the learning status of students in the new pilot mathematics program will be used by School Board members to access the status and content of the new curriculum. While this report can not derive predicative measure on how the new curriculum will fare when fully implemented, it can show the current spread of information.

Starting from the summary tables of schools and the comparison between the number of sessions (Figures 2.1-3.1), we can see a heavy cluster of data points around School A and B. However, when we take the average of the number of students at each school attending the new pilot math class (figure 3.1) , the data shows that the highest number of students per class attend high school E. Schools C and D have an overall over number of classes offered and the lowest number of students enrolled in the new math class. The results between the pattern of the number of students at each school is similar to the pattern for the distribution of the number of sessions at each school (Figure 4.1).

A boxplot that displays the general spread of student progress at all five schools showed that more students reported being behind in the new pilot-math class than any other category. Additionally, no students were reported being Very Ahead in the course, and only about 10 finished the course. The boxplot also shows outliers in Completed and More Behind.

Finally, the scatterplots (Figures 5.1-9.1) results showed that regardless of school, the highest number of students reported being behind in the new math class. Overall the distribution shows that more students report being behind, more behind, or very behind than being very ahead, middling, or completed with the new pilot math course. It is important to note that the plot for Schools A and B look different than those for Schools C, D, and E because of the lack of sessions at those schools. However, the scatterplot method is another method to look at spread of the data. From the graphs we can better detect outliers within each School. For example, schools A, B, and D have a clear outlier in the Behind variable. School C has a clear outlier in the Completed category.

Conclusion:

In summary, the school board can see from this data that the students who are taking the pilot math class this semester are struggling to stay with the class. This data was shown using a bar plot that compared the different learning statuses of students. A bar chart was also utilized twice to show, first, to display between schools and the number of sessions held; and secondly, when comparing the schools to the number of students enrolled in the new pilot math class. These visuals will help School Board Members see the spread of classes and where they are being held and the number of students enrolled in the new pilot-classes. Using this information, board members can make adequate suggestions, like choosing to focus on the schools with the highest number of students in the new program or requiring schools to cap each session at 15 students, so to build more individualized lessons and help.

Additionally, four scatterplots were created to group the results of the learning status by school. This will show School Board Members where to students stand within their individual schools.

It is recommended that the pilot program continue, with a decrease in the number of lessons taught, for a few semesters for a more representative data analysis. This way, variables can be plotted over time to better see what adjustments to the curriculum work and what does not.