The Relationship Between STEM Attitudes and Career Interests in Middle School

Introduction

Researchers use the term, STEM pipeline, to demonstrate "that the rates at which students currently enter STEM professions will be inadequate to keep up with the demands of our economy" (Cohen et al., 2013, p. 12). In order to build and sustain student interest in STEM disciplines, it is important first to ascertain the existing trends in their career choices and attitudes including factors influencing them. While most research focuses on students' high school or post-secondary experiences, performance, and interests as predictors of their STEM career pathways (e.g., Dewitt et al., 2011; Maltese & Tai, 2011; Hinojosa et al., 2016), new investigations reveal that the journey toward choosing a career starts earlier than that (Wiebe et al., 2018). While lacking a detailed understanding of all the career pathways in STEM, students in early grades are capable of differentiating and showing interest in STEM careers in broader ways (Maltese & Tai, 2011). For example, in 2006, Tai et al. analyzed the data collected for the 1988 US National Educational Longitudinal Study and showed that students who showed interest in a science-related career by the age of 14 were 3.4 times more likely to pursue a degree in physical sciences and engineering than those without similar interests. The correlation was even stronger for students with high mathematical capabilities, with 51% earning a STEM degree.

In order to address the gap in literature regarding students' career preferences in early years, this study aims to examine the relationship between students' attitudes toward different STEM disciplines (Science, Technology, Mathematics, and Engineering) and their career choices using the Student Attitudes Toward STEM Surveys [S-STEM] (Corn et al., 2012).

Research Question

What is the relationship between attitudes in STEM academic areas and STEM career interests in middle school?

Data

The data for the study was collected from 8 middle schools in Indiana. 270 students took the Student Attitudes Toward STEM (S-STEM) survey to record their responses. In the survey, there were three constructs to measure student attitudes toward the four primary STEM subjects: science (9 items), mathematics (8 items), and engineering/technology (9 items). For each discipline, a five-point Likert scale was utilized -- from strongly disagree to strongly agree to ask students for their responses. There was an additional construct that measured students' attitudes toward 21st century

skills. Additionally, to address external factors that may affect student interest in STEM, there were questions about whether or not students know adults working in STEM fields.

The S-STEM Survey measures student interest in twelve STEM career pathways: physics, environmental work, biology and zoology, veterinary work, mathematics, medicine, computer science, medical science, chemistry, energy, and engineering. A cluster analysis was run on student responses to these careers, and they were put into two clusters.

X Variable:

The numeric variables are the attitude scores of students in

1. Math
2. Science
3. Engineering & Tech
4. 21st-century skills

The categorical variables are

1. Do you plan to go to college?
2. Do you know any adults who work as scientists?
3. Do you know any adults who work as engineers?
4. Do you know any adults who work as mathematicians?
5. Do you know any adults who work as technologists?

Each categorical variable has three possible answers.

- Yes = 1
- No = 2
- Maybe = 0

Y Variable:

The two clusters of STEM careers were the Y variables:

- **Cluster 1**: Average scores in physics, math, computer science, chemistry, engineering, energy
- **Cluster 2**: Average scores in biology & zoology, environment work, veterinary work, medicine, medical science.

From the distribution of careers in the two clusters, it can be said that cluster 2 is predominated by biological science disciplines, whereas cluster 1 comprises mathematically rigorous disciplines.

Data Analysis:

The relationship between attitudes towards STEM disciplines and STEM career preferences was assessed using multiple linear regression in R with career cluster score being the dependent/Y variable and discipline attitude scoring the X variable.

Individual Scatter plot and Regression

The first step of the analysis was to plot a scatter plot with each of the 4 numeric independent variables against the 2 dependent variables to see if there were any obvious linear trends. It was then followed by running a regression with individual x and y variables.

Math

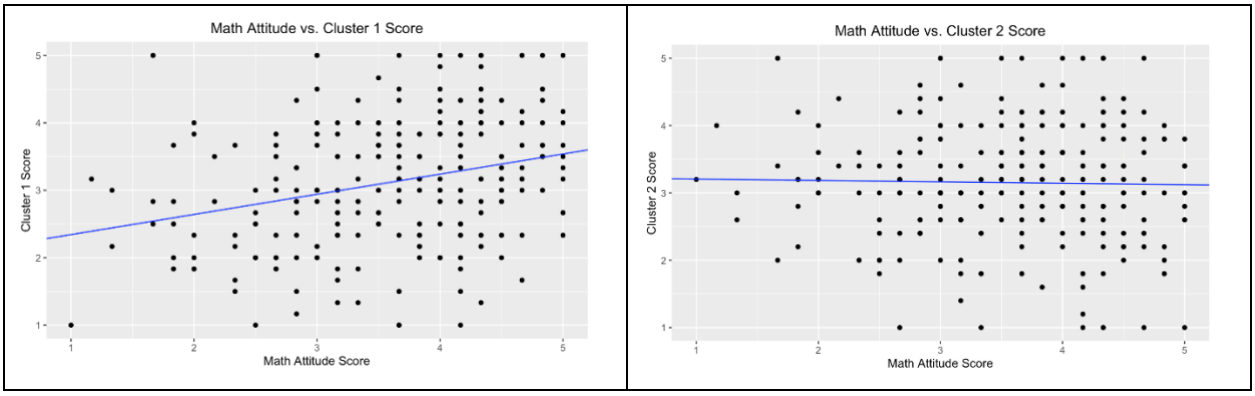


Figure 1 Scatter plots for variable math

The scatter plot for math attitudes vs. cluster 1 (Figure 1) shows a linear trend, however, the same cannot be said for the plot with cluster 2 scores. The regression outputs shown in Figure 2 confirm that math attitudes have a significant relationship with cluster 1 indicated by a p-value of less than 0.5 and math attitudes do not have a significant relationship with cluster 2.

```
Call:
lm(formula = cluster1 ~ math)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28907 -0.66455  0.01169  0.65154  2.45807

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.04384    0.23772   8.598 7.20e-16 ***
math         0.29886    0.06481   4.612 6.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9154 on 265 degrees of freedom
Multiple R-squared:  0.07429,   Adjusted R-squared:  0.0708
F-statistic: 21.27 on 1 and 265 DF,  p-value: 6.223e-06
```

```
Call:
lm(formula = cluster2 ~ math)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17178 -0.55242 -0.00699  0.63174  1.87047

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.22812    0.22824   14.14   <2e-16 ***
math        -0.02113    0.06222   -0.34    0.734
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8789 on 265 degrees of freedom
Multiple R-squared:  0.0004348, Adjusted R-squared:  -0.003337
F-statistic: 0.1153 on 1 and 265 DF,  p-value: 0.7345
```

Figure 2 Regression outputs for variable math

Science

Both the scatterplots in Figure 3 show that there exists a positive correlation between science attitude and the two clusters. Similarly, the simple linear regression outputs (Figure 4) for both the clusters show a significant positive correlation between science and the two clusters.

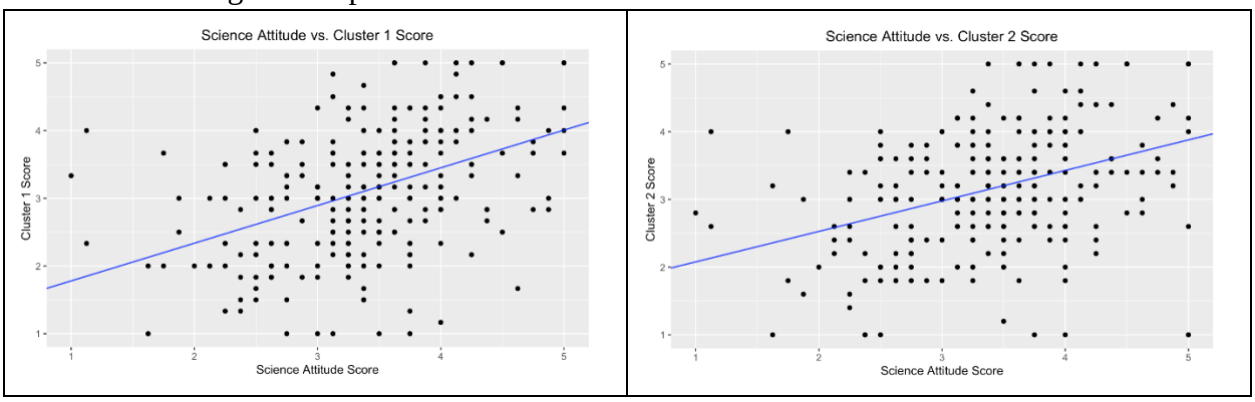


Figure 3 Scatter plots for variable science

```
Call:
lm(formula = cluster1 ~ science)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31075 -0.61662 -0.00488  0.55739  2.15090

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.22268    0.24296   5.032 8.95e-07 ***
science      0.55682    0.07003   7.952 5.35e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8549 on 265 degrees of freedom
Multiple R-squared:  0.1926,    Adjusted R-squared:  0.1896
F-statistic: 63.23 on 1 and 265 DF,  p-value: 5.345e-14
```

```
Call:
lm(formula = cluster2 ~ science)

Residuals:
     Min       1Q   Median       3Q      Max
-2.87917 -0.51589  0.02208  0.55310  1.86701

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.62603    0.23064   7.050 1.55e-11 ***
science      0.45063    0.06648   6.779 7.84e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8116 on 265 degrees of freedom
Multiple R-squared:  0.1478,    Adjusted R-squared:  0.1446
F-statistic: 45.95 on 1 and 265 DF,  p-value: 7.837e-11
```

Figure 4 Regression outputs for variable science

Engineering and Technology

The scatter plots (Figure 5) show that engineering and technology attitude score does indeed have a correlation with cluster1 and cluster2. The strength of the relationship is stronger for cluster 1 than cluster 2. This can also be confirmed from the greater magnitude of the coefficient in the simple linear regression output for cluster1 than cluster2 (Figure 6).

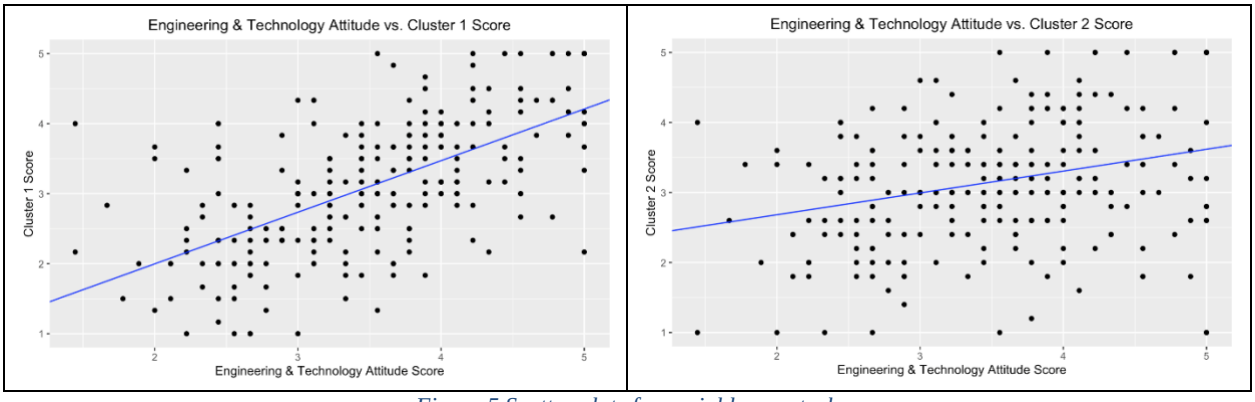


Figure 5 Scatter plots for variable eng_tech

```
lm(formula = cluster1 ~ eng_tech)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04268 -0.47643  0.00507  0.44166  2.41177

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.52340    0.20480   2.556   0.0112 *
eng_tech     0.73719    0.05692  12.951   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7446 on 265 degrees of freedom
Multiple R-squared:  0.3876,    Adjusted R-squared:  0.3853
F-statistic: 167.7 on 1 and 265 DF,  p-value: < 2.2e-16
```

```
lm(formula = cluster2 ~ eng_tech)

Residuals:
     Min       1Q   Median       3Q      Max
-2.61832 -0.52910  0.00556  0.64959  1.83226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.05861    0.23177   8.882  < 2e-16 ***
eng_tech     0.31194    0.06442   4.843 2.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8426 on 265 degrees of freedom
Multiple R-squared:  0.0813,    Adjusted R-squared:  0.07783
F-statistic: 23.45 on 1 and 265 DF,  p-value: 2.182e-06
```

Figure 6 Regression outputs for variable eng_tech

21st Century Skills

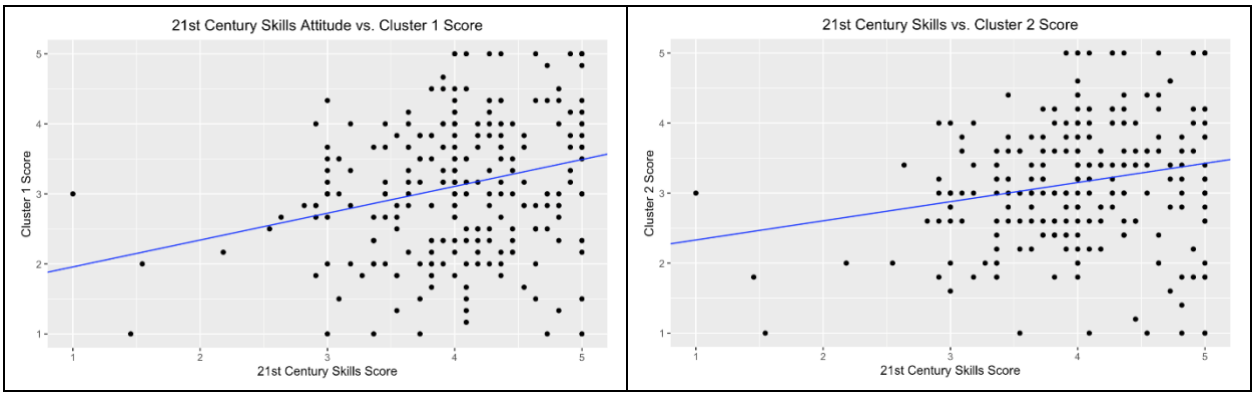


Figure 7 Scatter plots for variable tf_cent

The scatter plots (Figure 7) show that the 21st-century skills variable has a strong correlation with both response variables. This is evidenced in the positive sign of the coefficients in both the regression outputs (Figure 8) and the fact that the variable is significant at an alpha value of 0.05.

```
lm(formula = cluster1 ~ tf_cent)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4916 -0.7088  0.0586  0.7137  1.8919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.57374    0.34282   4.591 6.83e-06 ***
tf_cent      0.38358    0.08449   4.540 8.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9165 on 265 degrees of freedom
Multiple R-squared:  0.07217,   Adjusted R-squared:  0.06867
F-statistic: 20.61 on 1 and 265 DF,  p-value: 8.536e-06
```

```
lm(formula = cluster2 ~ tf_cent)

Residuals:
     Min       1Q   Median       3Q      Max
-2.42578 -0.55154  0.04803  0.59825  1.87292

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.05674    0.32168   6.394 7.27e-10 ***
tf_cent      0.27381    0.07928   3.454 0.000643 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.86 on 265 degrees of freedom
Multiple R-squared:  0.04308,   Adjusted R-squared:  0.03947
F-statistic: 11.93 on 1 and 265 DF,  p-value: 0.0006433
```

Figure 8 Regression outputs for variable tf_cent

Categorical Variable

Scatter plots are not applicable to categorical variables. One possible alternative is to look at the box plot of the response variable at each of the levels of the categorical variable. Similarly, t-test and the corresponding p-value are not appropriate for categorical variables. Therefore, an F test is used to check for significance of the categorical variable.
Only those variables that were selected by the model selection procedure are discussed in detail. The rest of the categorical variables showed no trend at all for both clusters and thus their outputs are not shown for brevity.

Scientist

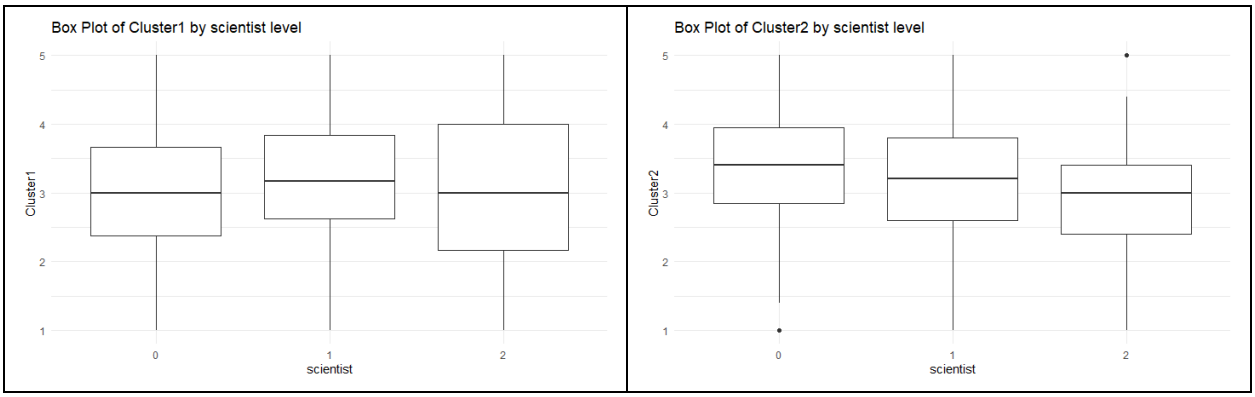


Figure 9 Box plots for categorical variable scientist

The box plots for the categorical variable scientist are shown in Figure 9. From the figure, we can see that there is no significant difference in the spread of the response variable cluster1 whereas, for cluster2, there seems to be a minor difference. This is confirmed using an ANOVA comparison as shown in Figure 10 where for cluster1 the categorical variable has no significance whereas for cluster2, it is significant at a confidence level of 90%.

```
```{r categorical F-test}
  null_model1 <- lm(cluster1 ~ 1, data = AC_Data);
  null_model2 <- lm(cluster2 ~ 1, data = AC_Data);

  model <- lm(cluster1 ~ factor(scientist), data = AC_Data);
  anova(null_model1, model);

  model <- lm(cluster2 ~ factor(scientist), data = AC_Data);
  anova(null_model2, model);
```

Analysis of Variance Table

Model 1: cluster1 ~ 1
Model 2: cluster1 ~ factor(scientist)
  Res.Df   RSS Df Sum of Sq   F Pr(>F)
1   266 239.90
2   264 237.41 2   2.4916 1.3853 0.252
Analysis of Variance Table

Model 1: cluster2 ~ 1
Model 2: cluster2 ~ factor(scientist)
  Res.Df   RSS Df Sum of Sq   F Pr(>F)
1   266 204.81
2   264 200.87 2   3.9367 2.587 0.07716 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10 ANOVA outputs for variable scientist

Technologist

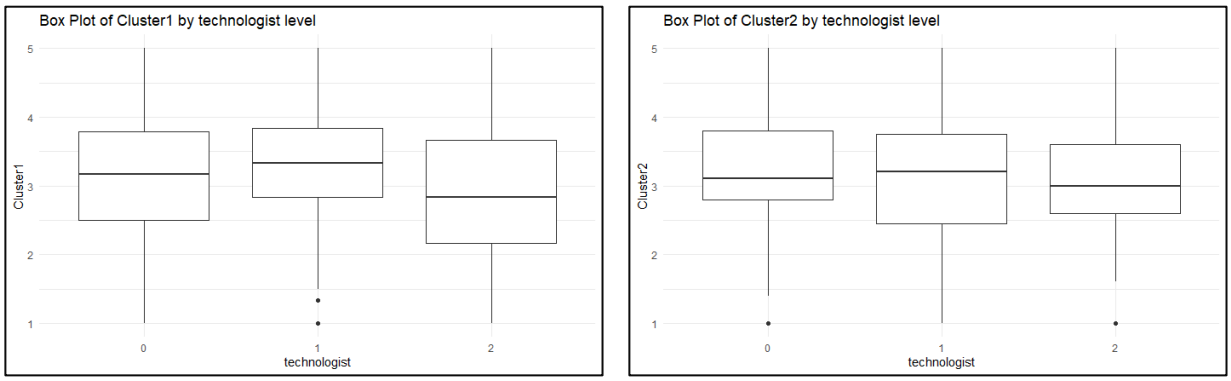


Figure 11 Box plots for categorical variable technologist

When examining the box plots in Figure 11, we see no pattern, indicating that the categorical variable technologist does not show any relationship with both clusters. The results of ANOVA (Figure 12) show that the variable technologist is significant at a level of 0.90 for cluster 1.

```r
```{r categorical F-test}
 null_model1 <- lm(cluster1 ~ 1, data = AC_Data);
 null_model2 <- lm(cluster2 ~ 1, data = AC_Data);

 model <- lm(cluster1 ~ factor(technologist), data = AC_Data);
 anova(null_model1, model);

 model <- lm(cluster2 ~ factor(technologist), data = AC_Data);
 anova(null_model2, model);
```
```

```
Analysis of Variance Table

Model 1: cluster1 ~ 1
Model 2: cluster1 ~ factor(technologist)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    266 239.90
2    264 235.35  2     4.542 2.5474 0.08021 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: cluster2 ~ 1
Model 2: cluster2 ~ factor(technologist)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    266 204.81
2    264 203.87  2   0.93311 0.6042 0.5473
```

Figure 12 ANOVA outputs for variable technologist

Model Selection

Model selection for Cluster 1

Once we have established that most of the independent variables have a significant linear relationship with at least one of the clusters, the next step is to identify a subset of variables that effectively explain the variation in the response variable. To do so, we use a stepwise model selection procedure. In particular, we used the procedure that does both the forward and backward selection. The result of this process (Figure 13) showed that the independent variables that explain the variation in cluster 1's distribution are engineering & technology, science, and mathematics.

```
#Model Selection
```{r}
#Model Selection with Cluster 1
datax <- data.frame(math, science, eng_tech, tf_cent, factor(college), factor(mathematican), factor(technologist),
factor(scientist), factor(engineer))
intercept1_only <- lm(cluster1~1, data = datax)
all_model1 <- lm(cluster1~., data = datax)
answer1 <- step(intercept1_only, direction = 'both', scope = formula(all_model1), trace = 0)
answer1$coefficients
```

 (Intercept)   eng_tech   science   math
  0.14618070 0.63647704 0.12786407 0.08338779
```

Figure 13 Model Selection output for Cluster 1

Regression for Cluster 1

Once the model selection process identified the best subset of variables, we ran a linear regression to confirm the significance of each variable. The output shown in Figure 14, verifies that engineering & technology has a significant linear relationship with cluster 1 score with $\beta = 0.6365$ at a significant level of 95%. Science also appears to be significant with $\beta = 0.12786$ for a significance level of 90%. While math is not significant, it is still being selected by the model selection. To verify if math is adding value to the model, we will run ANOVA with and without math to compare the SSE values.

```
lm(formula = cluster1 ~ eng_tech + science + math)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91671 -0.44696 -0.00137  0.43305  2.62384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14618    0.26283   0.556   0.5786
eng_tech     0.63648    0.07295   8.724 3.12e-16 ***
science      0.12786    0.07569   1.689   0.0924 .
math         0.08339    0.05542   1.505   0.1336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.74 on 263 degrees of freedom
Multiple R-squared:  0.3996,    Adjusted R-squared:  0.3928
F-statistic: 58.35 on 3 and 263 DF,  p-value: < 2.2e-16
```

Figure 14: Regression output for best subset of explanatory variables as provided by stepwise procedure for cluster 1

From the results shown in Figure 15, it appears that there is no major difference in the SSE value with or without math, suggesting that it does not add any more value to explaining the variation in the model. Hence, we will drop math and only keep engineering & technology and science for cluster 1.

```
Analysis of Variance Table

Response: cluster1
           Df  Sum Sq Mean Sq  F value  Pr(>F)
eng_tech    1  92.987  92.987 169.7948 < 2e-16 ***
science     1   1.641   1.641   2.9967 0.08461 .
math        1   1.240   1.240   2.2640 0.13361
Residuals 263 144.030   0.548
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

```
Analysis of Variance Table

Response: cluster1
           Df  Sum Sq Mean Sq  F value  Pr(>F)
eng_tech    1  92.987  92.987 168.9857 < 2e-16 ***
science     1   1.641   1.641   2.9824 0.08534 .
Residuals 264 145.269   0.550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Figure 15: ANOVA output for with and without variable Math for cluster 1

Hence, the beta estimators for the final model for Cluster 1 are,

```
Call:
lm(formula = cluster1 ~ eng_tech + science)

Residuals:
     Min       1Q   Median       3Q      Max
-1.98089 -0.43870 -0.01643  0.42766  2.55723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.33627    0.23102   1.456   0.1467
eng_tech     0.66402    0.07079   9.380   <2e-16 ***
science      0.13099    0.07585   1.727   0.0853 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7418 on 264 degrees of freedom
Multiple R-squared:  0.3945,    Adjusted R-squared:  0.3899
F-statistic: 85.98 on 2 and 264 DF,  p-value: < 2.2e-16
```

Figure 16: Regression output for final model of cluster 1

Model Selection for Cluster 2

Similarly, step-wise model selection was used to determine the best subset of variables that explained the variation exhibited by Cluster 2. The results of this procedure (shown in Figure 17) indicate that the independent variables that best explain the variation in cluster 2's distribution are science, math, engineering & technology, scientist, and technologist.

```{r}
#Model Selection with Cluster 2
datax <- data.frame(math, science, eng_tech, tf_cent, factor(college), factor(mathematican), factor(technologist),
factor(scientist), factor(engineer))
intercept_only <- lm(cluster2~1, data = datax)
all_model2 <- lm(cluster2~., data = datax)
answer2 <- step(intercept_only, direction = 'both', scope = formula(all_model2), trace = 0)
answer2$coefficients
```

```
          (Intercept)               science                  math              eng_tech     factor.scientist.1
            1.8982735             0.3871923            -0.1440238             0.1697908            -0.1051040
factor.scientist.2 factor.technologist.1 factor.technologist.2
           -0.3334797            -0.1220322             0.1685707
```

Figure 17: Model Selection Output for Cluster 2

Regression model for Cluster 2

To confirm if the individual explanatory variables were statistically significant, a linear regression was performed. The output is shown in Figure 18. We can see that the continuous variables, science, math, and eng_tech are significant at $\alpha = 0.95$. One of the levels of the categorical value scientist is significant (level 2, indicating a 'no' response). The categorical variable technologist is not significant even at $\alpha = 0.90$ but it is still suggested by model selection. To confirm the contribution of technologist in the model, we will run ANOVA.

```
lm(formula = cluster2 ~ science + math + eng_tech + factor(scientist) +
    factor(technologist))

Residuals:
     Min       1Q   Median       3Q      Max
-2.73593 -0.50917 -0.01512  0.52122  1.86587

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.89827    0.30212   6.283  1.4e-09 ***
science               0.38719    0.08186   4.730  3.7e-06 ***
math                 -0.14402    0.06023  -2.391   0.0175 *
eng_tech              0.16979    0.07981   2.127   0.0343 *
factor(scientist)1   -0.10510    0.12558  -0.837   0.4034
factor(scientist)2   -0.33348    0.13492  -2.472   0.0141 *
factor(technologist)1 -0.12203    0.12378  -0.986   0.3251
factor(technologist)2  0.16857    0.13518   1.247   0.2135
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7967 on 259 degrees of freedom
Multiple R-squared:  0.1973,    Adjusted R-squared:  0.1756
F-statistic: 9.096 on 7 and 259 DF,  p-value: 4.828e-10
```

Figure 18: Regression output for best subset of explanatory variables as provided by the stepwise procedure for cluster 2

From the ANOVA table in Figure 19, we find contrasting results. While technologist happen to become significant at $\alpha = 0.90$, scientist become insignificant. This is contrary to the results found in linear regression shown in Figure 18.

```
Analysis of Variance Table

Response: cluster2
                     Df  Sum Sq Mean Sq F value    Pr(>F)
science               1  30.267 30.2665 47.6848 3.856e-11 ***
math                  1   2.308  2.3077  3.6358   0.05766 .
eng_tech              1   1.986  1.9862  3.1292   0.07808 .
factor(scientist)     2   2.568  1.2838  2.0226   0.13440
factor(technologist)  2   3.285  1.6424  2.5876   0.07714 .
Residuals           259 164.393  0.6347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 19: ANOVA output for best subset of explanatory variables as provided by the stepwise procedure for cluster 2

The t-test results shown in Figure 18 show that level 2 response (‘no’ response) for a categorical variable scientist was significant. This indicates that there is a statistically significant difference in the mean response between level 2 and the base case (level 0, or ‘maybe’ response). This prompted us to check if level 2 is significantly different than combined level 0 and level 1.

The regression analysis was performed with new variables technologist and scientist, that only had two levels:

- ‘2’ for ‘no’ response
- ‘1’ for ‘yes’ or ‘maybe’ response

The result of the regression analysis is shown in Figure 20. The new variables with changed levels are all significant now. The result of ANOVA also confirms the same.

```
Call:
lm(formula = cluster2 ~ science + math + eng_tech + factor(scientist) +
    factor(technologist))

Residuals:
     Min       1Q   Median       3Q      Max
-2.82186 -0.51090  0.00007  0.55504  1.78669

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.83039    0.29854   6.131 3.21e-09 ***
science               0.38305    0.08181   4.682 4.57e-06 ***
math                 -0.15205    0.05996  -2.536   0.0118 *
eng_tech              0.16730    0.07977   2.097   0.0369 *
factor(scientist)2   -0.27629    0.11273  -2.451   0.0149 *
factor(technologist)2  0.22319    0.11503   1.940   0.0534 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7969 on 261 degrees of freedom
Multiple R-squared:  0.1908,    Adjusted R-squared:  0.1753
F-statistic: 12.31 on 5 and 261 DF,  p-value: 9.989e-11
```

Figure 20: Regression output after change of levels for scientist and technologist variables

This was further confirmed by ANOVA comparison of null and alternative hypothesis without and with technologist respectively. The results in Figure 20 show that the full model with technologist does add value to the model. Similarly, adding scientist to the model significantly explains the variations in fitted Y, as shown in Figure 21.

```{r}
Newest_model_cluster2 <- lm(cluster2~science+math+eng_tech+factor(scientist)+factor(technologist))
summary(Newest_model_cluster2)
Newest_model_cluster2.Null <- lm(cluster2~science+math+eng_tech+factor(scientist))

#Comparing Full and Reduced Model
anova(Newest_model_cluster2.Null,Newest_model_cluster2)
```

```
Analysis of Variance Table

Model 1: cluster2 ~ science + math + eng_tech + factor(scientist)
Model 2: cluster2 ~ science + math + eng_tech + factor(scientist) + factor(technologist)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    262 168.12
2    261 165.73  1    2.3903 3.7643 0.05343 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 21:ANOVA comparison to test significance of the variable technologist after level change

```{r}
Newest_model_cluster2 <- lm(cluster2~science+math+eng_tech+factor(technologist)+factor(scientist))
#summary(Newest_model_cluster2)
Newest_model_cluster2.Null <- lm(cluster2~science+math+eng_tech+factor(technologist))

#Comparing Full and Reduced Model
anova(Newest_model_cluster2.Null,Newest_model_cluster2)
```

```
Analysis of Variance Table

Model 1: cluster2 ~ science + math + eng_tech + factor(technologist)
Model 2: cluster2 ~ science + math + eng_tech + factor(technologist) +
    factor(scientist)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    262 169.54
2    261 165.73  1    3.8147 6.0076 0.0149 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 22: ANOVA comparison to test significance of the variable scientist after level change

Diagnostic Tests

Multiple linear regression (MLR) has fundamental assumptions which if violated, invalidate the model. In the previous section, the best set of explanatory variables was identified for each of the clusters. In this section, diagnostic tests are performed to assess if any of the fundamental assumptions are violated.

1. Normality

Shapiro-Wilk Test: It formally tests the normality of residuals using the Shapiro-Wilk test. The null hypothesis is that the residuals are normally distributed. For both cluster 1 and cluster 2, we fail to reject the null hypothesis at a significance level of $\alpha = 0.95$. Therefore, we conclude that the residuals are normally distributed.

Normal Q-Q Plot: It checks if the residuals follow a normal distribution. A Q-Q plot compares the quantiles of the residuals to the quantiles of a normal distribution. For cluster 1 (Figure 23), there is no significant deviation from normality. For cluster 2 (Figure 24), the QQ plot shows a slight deviation in the lower tail, but Shapiro-Wilk's test shows that there is no issue with normality, thus no remedial measure was applied.

```
        Shapiro-Wilk normality test

data:  residuals(Newest_model_cluster1)
W = 0.99403, p-value = 0.3736
```

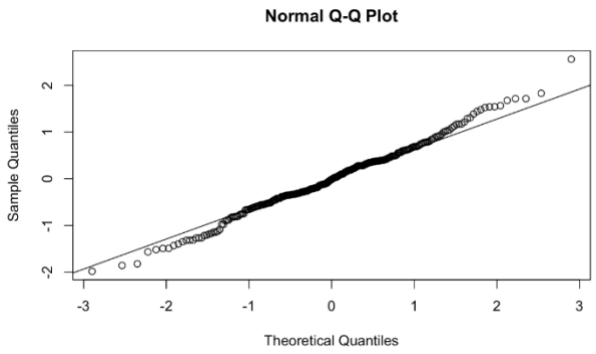


Figure 23: Tests for normality assumption – Cluster 1

```
        Shapiro-Wilk normality test

data:  residuals(Newest_model_cluster2)
W = 0.99158, p-value = 0.1308
```

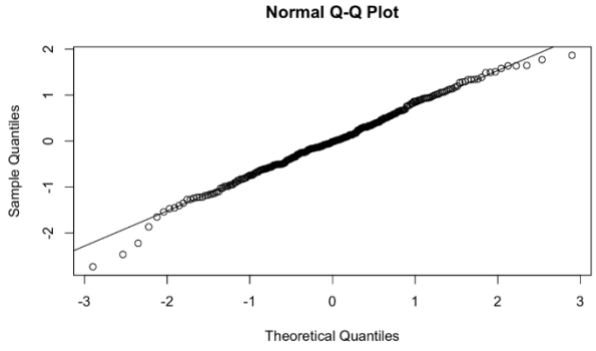
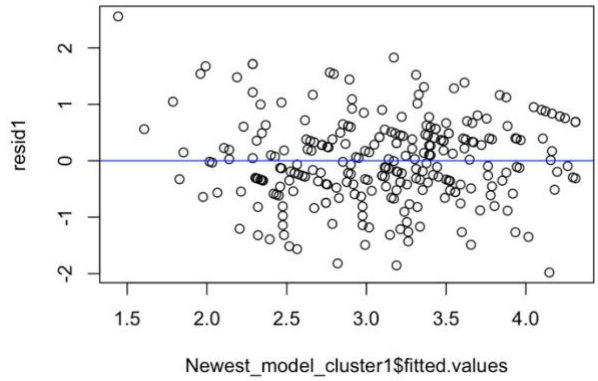


Figure 24: Tests for normality assumption – Cluster 2

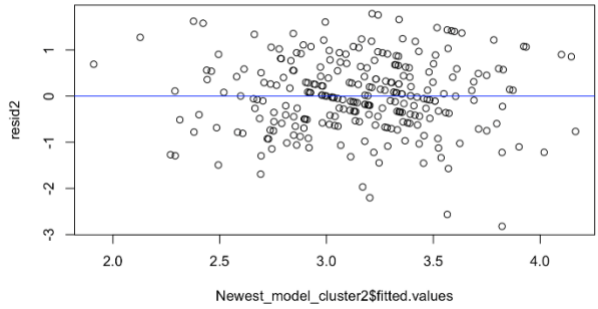
2. Constant Variance

To test for constant variance, we can use residual vs. fitted value plots or the Breusch Pagan test.
Residuals vs. Fitted Values Plot: This plot is used to visually inspect for any patterns between the residuals and the fitted values. The absence of a pattern indicates constant variance. From Figure 25 and Figure 26 the random patterns of the residual plots indicate that variance is constant.
Breusch-Pagan Test or White Test: Formally test for heteroscedasticity in residuals. The null hypothesis for the Breusch Pagan test is that there is no heteroscedasticity. For cluster 1, we fail to reject the null hypothesis as p-value is greater than 0.05. Similarly, we fail to reject the null hypothesis for cluster 2.




```
        studentized Breusch-Pagan test

data:  Newest_model_cluster1
BP = 5.342, df = 2, p-value = 0.06918
```

Figure 25: Tests for constant variance assumption – Cluster 1




```
        studentized Breusch-Pagan test

data:  Newest_model_cluster2
BP = 10.752, df = 5, p-value = 0.05653
```

Figure 26: Tests for constant variance assumption – Cluster 2

3. Multicollinearity between Independent Variables

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to isolate the individual effects of each variable on the dependent variable. The multicollinearity plots (Figure 27) and the correlation matrix (Figure 28) do not show any unusual correlation between the independent variables. However, to verify further VIF will be used.

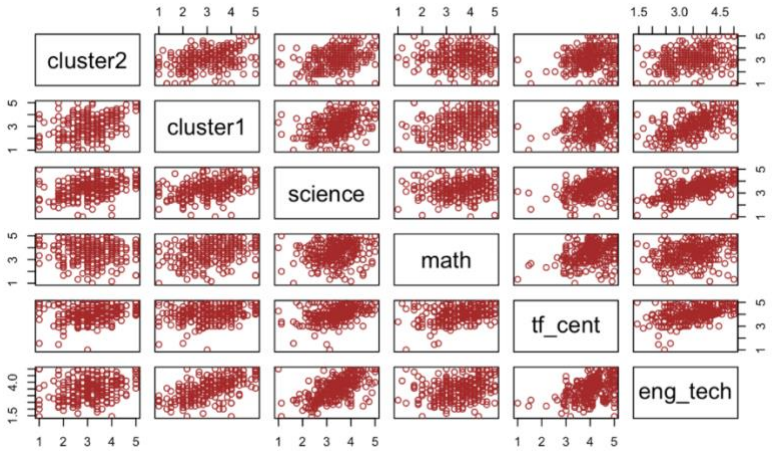


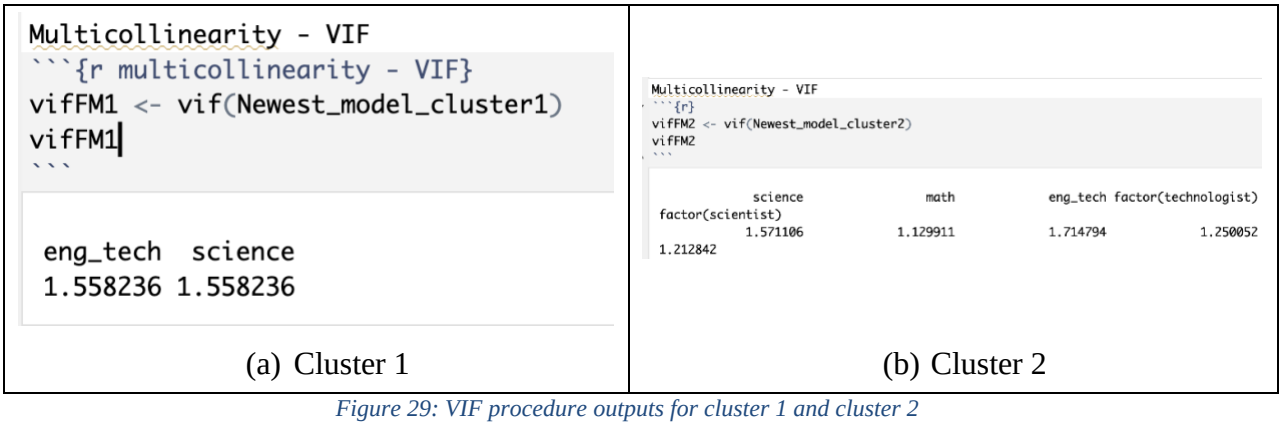
Figure 27: Multicollinearity plots

	Cluster 1	Cluster 2	math	science	engTech	tfcent
Cluster 1	1.0000	0.4772	0.2726	0.4389	0.6226	0.2686
Cluster 2	0.4772	1.0000	0.0209	0.3844	0.2851	0.2075
math	0.2726	0.0209	1.0000	0.2154	0.3252	0.3655
science	0.4389	0.3844	0.2154	1.0000	0.5985	0.4035
engTech	0.6226	0.2851	0.3252	0.5985	1.0000	0.4619
tfcent	0.2686	0.2075	0.3655	0.4035	0.4619	1.0000

Figure 28: Correlation Matrix

VIF for the new models

VIF is a measure used to quantify the severity of multicollinearity in a multiple regression model. Typically, a VIF value of 1 indicates no multicollinearity, and as the VIF value increases, the severity of multicollinearity increases. A common rule of thumb is that VIF values greater than 5 or 10 are a cause for concern. The results of VIF for both clusters 1 and 2 show that there is no multicollinearity issue in the data.



(a) Cluster 1 (b) Cluster 2

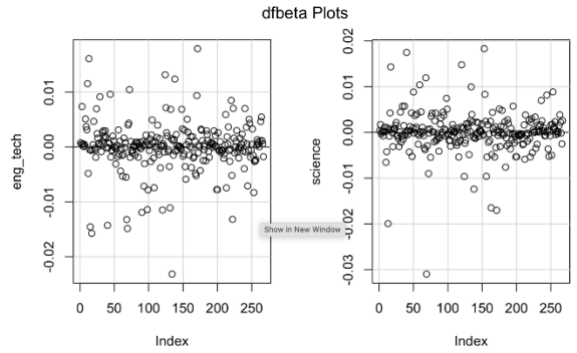
Figure 29: VIF procedure outputs for cluster 1 and cluster 2

4. Influential Points and Outliers

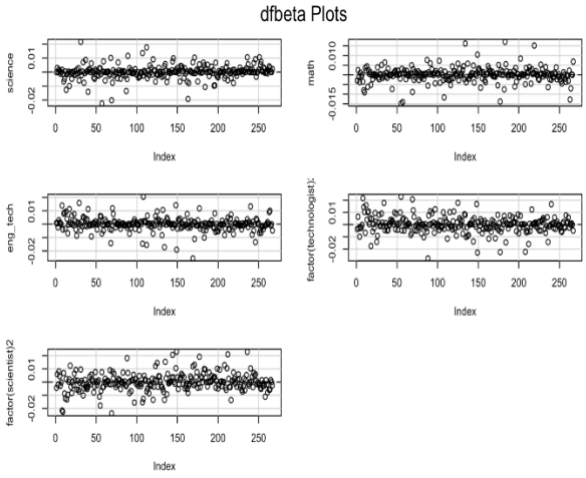
Influential points and outliers refer to specific observations that can have a significant impact on the results of the regression analysis. These points can influence the estimation of the regression coefficients, the overall model fit, and the predictions. It's important to identify and understand influential points and outliers to assess the robustness of your regression model.

Influential Points:

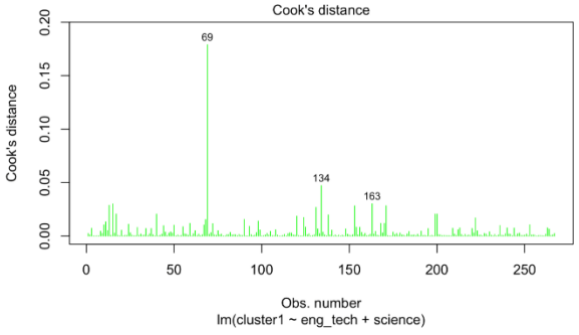
Influential points are observations that, if removed from the dataset, could substantially change the regression estimates. Cook's Distance is used to quantify the influence of individual observations on the fitted values. High Cook's Distance values indicate influential points. Given that the input data ranges between 1 and 5, the lack of influential observations for both clusters are not surprising (Figure 30). DFBETA is used to quantify the influence of individual observations on the regression coefficients. The DFBETA plots in Figure 30 show that there are no points with a value greater than 1, implying that there are no points that wield a significant influence on the estimated regression coefficients.



Influential points on regression coefficients

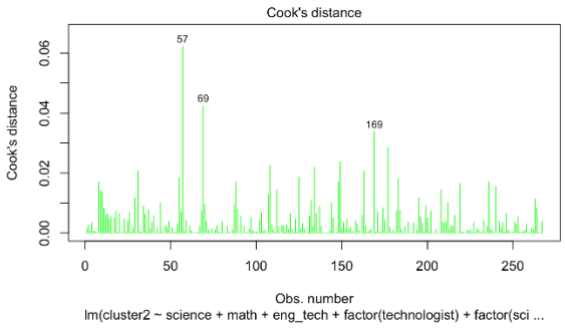


Influential points on regression coefficients



Influential points on fitted values

(a) Cluster 1



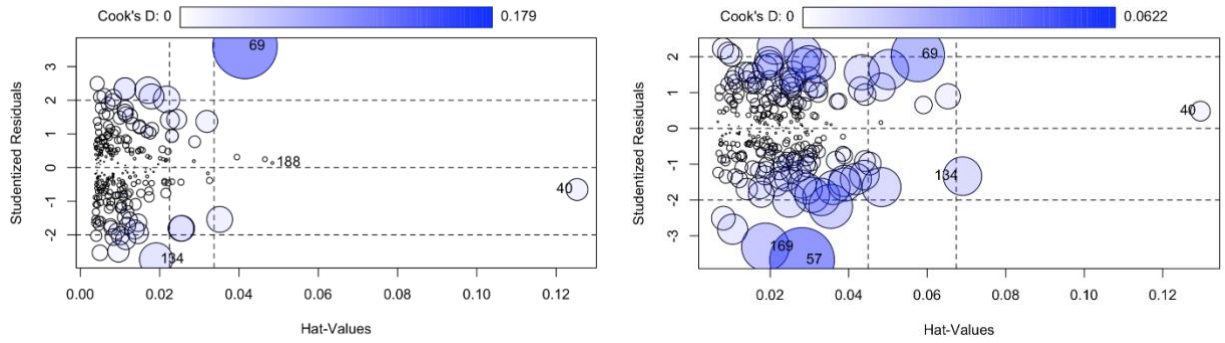
Influential points on fitted values

(b) Cluster 2

Figure 30: Diagnostics for influential points

Outliers:

Outliers are data points that deviate significantly from the overall pattern of the data. There are studentized residuals greater than ± 2 indicating the presence of outliers on Y. Hat values are greater than $\frac{2p}{n}$ which is 0.022 and 0.044 for cluster 1 and cluster 2 respectively. This indicated that there are outliers in X.



(a) Cluster 1 (b) Cluster 2

Figure 31: Outliers

Remedial Measure(s)

Robust Regression

Since the models for both cluster 1 and cluster 2 had outliers, robust regression was used to trim the influence of these extreme observations on estimations. The weight function chosen here is bisquare for dampening the influence of outliers based on their residuals. The R output is shown in Figure 32.

The residual standard error in our model with cluster 1 before robust weighting is 0.7418 (Figure 16). This value reduces to 0.6446 after robust weighting as seen in plot (a) of Figure 32 indicating that Robust weighting provided a better fit. Similarly, we can compare the residual standard error for cluster 2 model before and after robust weighting. Its value has increased from 0.7969 to 0.8035. Although there is a marginal increase observed, it is so slight that it likely has minimal impact on the overall analysis.

```
Call: rlm(formula = cluster1 ~ science + eng_tech, psi = psi.bisquare)
Residuals:
      Min        1Q    Median        3Q       Max
-2.071785 -0.431037 -0.006489  0.437863  2.784340

Coefficients:
            Value   Std. Error t value
(Intercept) -0.0225  0.2301    -0.0977
science      0.1728  0.0755     2.2884
eng_tech     0.7226  0.0705    10.2500

Residual standard error: 0.6446 on 264 degrees of freedom
Analysis of Variance Table

Response: cluster1
          Df  Sum Sq Mean Sq F value Pr(>F)
science    1  50.338  50.338
eng_tech   1  50.509  50.509
Residuals    146.621
```

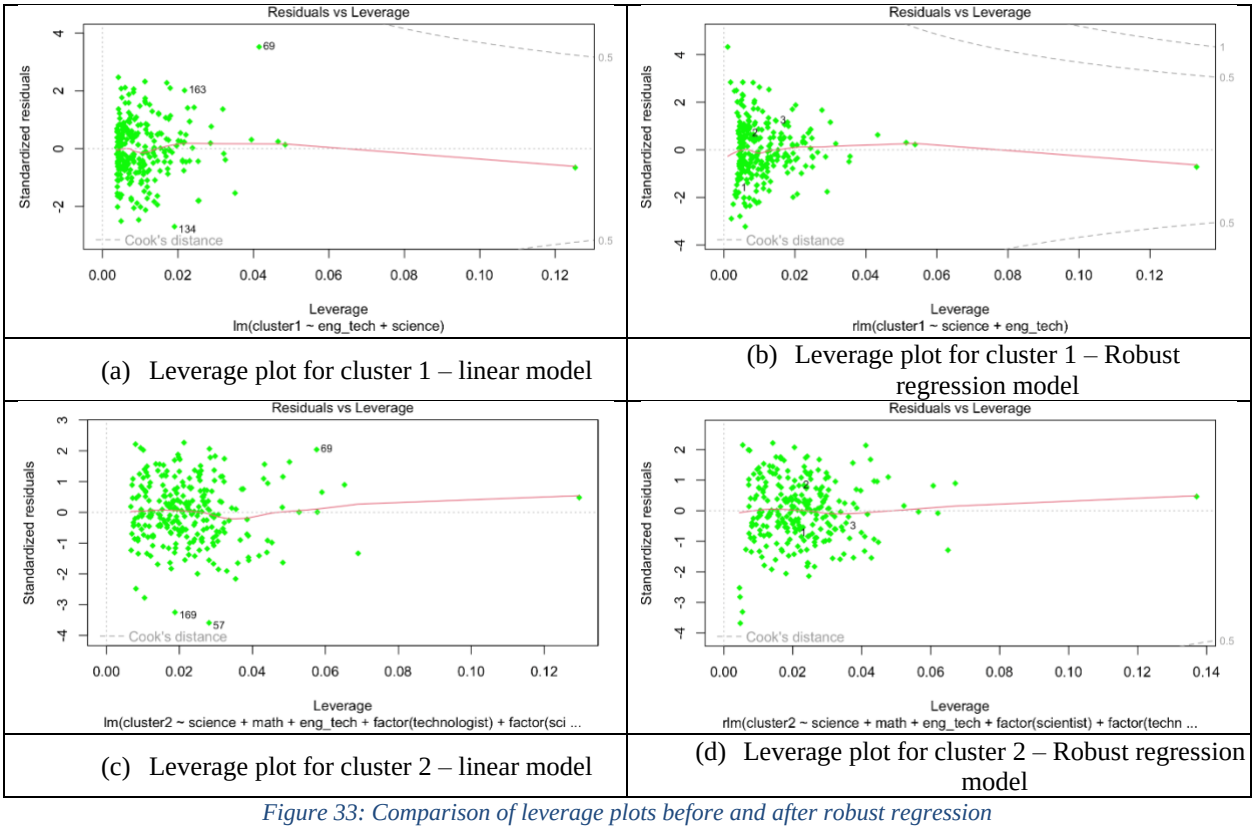
(a) Cluster 1

```
Call: rlm(formula = cluster2 ~ science + math + eng_tech + factor(scientist) +
    factor(technologist), psi = psi.bisquare)
Residuals:
     Min       1Q   Median       3Q      Max
-2.95146 -0.50846 -0.02779  0.54484  1.76538

Coefficients:
                    Value   Std. Error t value
(Intercept)          1.6423  0.3012     5.4522
science              0.4064  0.0825     4.9230
math                -0.1290  0.0605    -2.1325
eng_tech             0.1845  0.0805     2.2919
factor(scientist)2  -0.3138  0.1137    -2.7586
factor(technologist)2  0.2132  0.1161     1.8366

Residual standard error: 0.8035 on 261 degrees of freedom
Analysis of Variance Table

Response: cluster2
                   Df  Sum Sq Mean Sq F value Pr(>F)
science             1  32.153  32.153
math                1   1.182   1.182
eng_tech            1   2.314   2.314
factor(scientist)   1   2.915   2.915
factor(technologist)  1   2.014   2.014
Residuals             166.382
```

(b) Cluster 2

Figure 32: Robust regression results

Leverage measures how far an independent variable value (or a set of values, in the case of multiple regression) deviates from its mean. High-leverage points are those with extreme predictor (X) values; they have more potential to influence the regression line. Points with high leverage can disproportionately affect the slope of the regression line, even if their residuals are small. Robust regression tries to reduce the leverage of points that have very high residuals. This can be evidently seen in the graphs shown in Figure 33.



(a) Leverage plot for cluster 1 – linear model

(b) Leverage plot for cluster 1 – Robust regression model

(c) Leverage plot for cluster 2 – linear model

(d) Leverage plot for cluster 2 – Robust regression model

Figure 33: Comparison of leverage plots before and after robust regression

Bootstrapping

Bootstrapping is a resampling technique in which samples are drawn with replacement from the observed data to create multiple bootstrap samples. Each of these samples is then used to estimate the variability, confidence intervals, or other statistical properties of a population parameter. The original estimates are the parameter estimates obtained from the initial fitting of the robust regression model to the original data while the bias estimate shows the average difference between the bootstrap estimates and the original estimate. The results from the bootstrapping procedure are shown in Figure 34.

In the case of cluster 1, a positive bias suggests that, on average, the bootstrap estimates are higher than the original estimate. Lastly, the non-zero std, error shows variability in bootstrap estimates.

In Cluster 2, the biases for each coefficient seem relatively small, and the signs (positive or negative) indicate the direction of the bias. The standard error of the bootstrap distribution reflects the variability of the bootstrap estimates around their mean.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = datax, statistic = boot.robustCoef, R = 1000)


Bootstrap Statistics :
       original     bias    std. error
t1* -0.02247644  3.1259043  0.29348193
t2*  0.17283989 -0.1696864  0.10504453
t3*  0.72255574 -0.7241793  0.09615308
```

t1*: intercept
t2*: coefficient for the variable 'science'
t3*: coefficient for the variable 'eng_tech'

(a) Cluster 1

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = datax, statistic = boot.robustCoef, R = 1000)


Bootstrap Statistics :
      original      bias    std. error
t1*  1.6422999  1.61919206  0.31551941
t2*  0.4063805 -0.40751341  0.09176964
t3* -0.1290223  0.12722531  0.06494435
t4*  0.1844730 -0.18550097  0.08590600
t5* -0.3137651 -0.02031315  0.01440853
t6*  0.2131637 -0.13287419  0.01392532
```

t1*: intercept.
t2*: coefficient for the variable 'science').
t3*: coefficient for the variable 'math'
t4*: coefficient for the variable ‘eng_tech’
t5*: coefficient for the variable 'scientist'
t6*: coefficient for the variable 'technologist'

(b) Cluster 2

Figure 34: bootstrapping procedure output

The confidence intervals for each of the betas obtained via bootstrapping are shown in Table 1.

Table 1 Confidence interval obtained from bootstrapping.

Model	Parameter	95th Percentile	
Cluster 1	Intercept	2.5317	3.6581
	Science	-0.1975	0.02051
	Engineering & Tech	-0.1971	0.1792
Cluster 2	Intercept	2.643	3.859
	Science	-0.1745	0.1906
	Math	-0.1259	0.1206
	Engineering & Tech	-0.1741	0.1692
	Scientist	-0.3637	-0.3041
	Technologist	0.0522	0.1086

DISCUSSION

This analysis provides deeper insight into the body of research concerning students' attitudes toward STEM disciplines and STEM career pathways in middles schools. The model selection method provided a productive starting point to shortlist significant variables, which were validated using ANOVA and further improved by running diagnostics. The linear models for the two clusters were found to be,

$$Cluster\ 1 = \beta_o + \beta_1\ Science + \beta_2\ Eng\&Tech$$
$$Cluster\ 1 = -0.0225 + 0.1728\ Science + 0.7226\ Eng\&Tech$$

$$Cluster\ 2 = \beta_o + \beta_1\ Science + \beta_2\ Math + \beta_3\ Eng\&Tech + \beta_4\ Scientist + \beta_5\ Technologist$$
$$Cluster\ 2 = 1.6423 + 0.4064\ Science - 0.1290\ Math + 0.1845\ Eng\&Tech - 0.3138\ Scientist + 0.2132\ Technologist$$

While Cluster 1 comprised of careers for physics, math, computer science, chemistry, engineering, and energy, Cluster 2 had careers in biology & zoology, environment work, veterinary work, medicine, and medical science. It is interesting to notice that despite comprising computationally rigorous career options, the final model for Cluster 1 did not have math attitude as a significant variable. This means there is a disconnect between math attitudes and opting for computationally rigorous careers. On the other hand, Cluster 2 was dominated by biology-related careers and had a negative estimator for math attitudes. This aligns with previous findings that students who are less inclined towards mathematics tend to choose non-mathematical career paths. Furthermore, both science and engineering & technology were found to be significant positive contributors to both clusters. While Engineering & Technology had a larger estimator value (0.7226) for Cluster 1, it had a comparatively smaller value (0.1845) for Cluster 2. On the contrary, Science had a rather smaller estimator value (0.1728) for Cluster 1 but a larger one (0.4064) for Cluster 2. Lastly, for Cluster 1, none of the categorical variables were found to be significant. However, for Cluster 2, it appeared that while having known a scientist had a negative impact, knowing of a technologist in the family had a positive impact on student choice towards the careers in the bin.

REFERNCES

Cohen, C., Patterson, D. G.,Kovarik, D. N., Chowning, J. T. (2013). *Fostering STEM career awareness: Emerging opportunities for teachers.* Washington State Kappan.

DeWitt, J., Osborne, J., Archer, L., Dillon, J., Willis, B., & Wong, B. (2011). Young children's aspiration in science: The unequivocal, the uncertain and the unthinkable. *International Journal of Science Education, 35*(6), 1037–1063. https://doi.org/10.1080/09500693.2011.608197

Hinojosa, T., Rapaport, A., Jaciw, A.P., & Zacamy, J. (2016). *Exploring the Foundations of the Future STEM Workforce: K-12 Indicators of Postsecondary STEM Success*. REL 2016-122.

Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. Students. Science Education, 95(5), 877–907

Wiebe, E., Unfried, A., & Faber, M. (2018). The Relationship of STEM Attitudes and Career Interest. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(10), em1580.