Unlocking the Key to Success: Factors Influencing Student Learning and Career Decisions in the STEM Field

Samriddhi Soni*, Sharon Qian**

* Human Development, Teachers College, ss6557@tc.columbia.edu
** International and Comparative Education, Teachers College, sq2233@tc.columbia.edu

Abstract

The impact of family background, reading habits, and STEM performance on student achievement and career aspirations has been the subject of much research. Researchers in this study employ multiple educational data mining techniques such as clustering, predictive analysis, and visualization to study factors and relationships on academic achievement and career choices. The findings suggest that family background, in particular, has a significant impact on children's academic achievement and opportunities for success. Reading habits and STEM performance have also been shown to be important determinants of student achievement and career aspirations. Improving performance in these areas can help increase the number of students pursuing STEM careers and promote competitiveness and economic prosperity. A holistic approach that considers the interplay of these factors is necessary to foster a STEM-capable workforce.

Keywords

STEM career opportunities, socio-economic factors, student learning outcomes, student reading habits.

1. INTRODUCTION

Education is a key determinant of students' future success. Recent years have seen a significant increase in the importance of Science, Technology, Engineering, and Mathematics (STEM) education in fostering the growth of a highly educated and technologically competent workforce. Therefore, it is essential to promote competitiveness and economic prosperity consciously through STEM education. In order to foster a STEM-capable workforce and emphasize STEM education, there is a need to understand the factors that influence student achievement and career aspirations.

Existing research has mostly focused on the impact of family background on eventual educational attainment. Nores and Barnett (2005) find that parental expectations and the home environment play an important role in mediating the relationship between these factors and child achievement. In specific, Welsh and Lippman (2009) investigate the same question by using data from the 2003 National Assessment of Educational Progress. The findings show that students whose parents have higher levels of education achieve higher test scores. Cascio and Schanzenbach (2015) continue on Nore and Barnett (2005)'s study by examining the relationship between family income and academic achievement. The authors find that children from low-income families are more likely to have lower academic achievement compared to their peers from higher-income families. Guided by a sociological perspective, Duncan and Murnane (2011) discuss the relationship between rising inequality, schools, and children's life chances. The authors argue that increasing inequality is threatening children's opportunities for success in school and in life. Notably, Ronfard and Guimond (2018) present a meta-analysis of the long-term effects of family income on children's academic achievement. They find that family income has a significant negative impact on children's academic achievement. Additionally, recent studies have shown that reading practices and STEM performance also play a significant role in predicting future success. Braun et al. (2005) examine the relationship between reading and science achievement among secondary school students in the United States. The authors find that students who spend time reading additional books have better performance in both reading and scientific subjects than those who do not. Fang and Wei (2008) present a meta-analysis of the impact of reading and writing on science achievement. The authors find that reading and writing have a significant positive effect on science achievement. The article by Ranjeeth et al. (2017) investigates the relationship between academic performance and outside-class habits of Indian secondary school students. The authors find that students who engage in extracurricular activities, such as reading and sports, have better academic performance compared to those who do not.

Among the literature available, some scholars have indicated that student academic performance can impact career aspirations among secondary students. One research article on the influence of student STEM performance on their future job choices was conducted by Chen et al. (2018). The study uses data from the Programme for International Student Assessment (PISA) to examine the relationship between students' performance in STEM subjects and their likelihood of pursuing STEM careers. The study suggests that improving student performance in STEM subjects could help increase the number of students pursuing STEM careers. In the same year, Schaller and Scott (2018) use data from the National Education Longitudinal Study of 1988 to investigate the impact of high school science and mathematics courses on students' postsecondary STEM field choices. The results indicate that students who take more advanced science and mathematics courses in high school are more likely to choose STEM fields of study in college. The study by Hayes et al. (2012) focuses on the role of science identity and science interest in the STEM career choice process. The authors found that students who had a stronger science identity and higher levels of science interest were more likely to choose a STEM career. Other related studies focusing on the influence of student learning habits on their future job choices exist in the field of learning analytics. Shariff et al. (2017) find that students who have strong learning habits are more likely to succeed in the learning analytics course. Additionally, Charfuelan et al (2018) expand Shariff et al. (2017)'s study and find a significant relationship between learning habits and student performance in a learning analytics course. Durmaz et al. (2017) look at how students' learning habits and their ability to control their own learning affect how well they perform in a learning analytics course. In accordance with Durmaz et al. (2017)'s research, Taspinar et al. (2018) conducted similar research focusing on student self-regulated learning. Both studies found that both learning habits and self-regulated learning were significant predictors of student success in the course. These articles provide valuable insights into the importance of learning habits for student success in the field of science education.

The main purpose of this research is to investigate contributing factors on student STEM achievement and career decisions. This study addresses three primary research questions based on the literature described above:

1. To what extent does a family's socioeconomic background influence student learning?
2. Do student's reading habits influence their educational achievements?
3. What is the relationship between students' STEM performance, learning habits and career choices?

By examining these three research questions, this study will provide valuable insights into the factors that impact student learning and career decisions in the field of education mining.

2. RESEARCH METHODOLOGY

2.1 Sample

The current research sample is based on datasets from students at 16 junior high schools in Ninh Binh Province, Vietnam, ranging in age from 11 to 15 years old. There are 2464 female students and 2451 male students from Grade 6 to Grade 9, for a total of 4915 students. Students in this study were classified as being of poor, medium, or rich socioeconomic status.

2.2 Data Source

The source of this study is drawn from the datasets on secondary school students' STEM performance and reading practices which are collected by Vuong et al. (2021). Following are the variables used in this study. "Student_ID" indicates the identification variable of students. The personal information category includes "Grade_level" which is students from Grade 6 to Grade 9 and "Sex" with male and female. "Quiz" represents the average of the marks scored in STEM subjects: Mathematics, physics, chemistry and biology. As for "Hobby", students were given six main kinds of activities, namely reading books, watching TV/listening to music, housework/farming, observing nature, interacting with friends/family members which are assigned to "a", "b", "c", "d", "e" and "f". "Career_goal" includes future occupations that students are interested in. Regarding the category of family background, "Career_father" and "Career_mother"means the occupation of the student's father and mother. "Career_father(factor)" and "Career_mother(factor)" indicate the occupation of the student's father which is factored into two factors namely: Labor and intellectual. In addition, "Edu_father" and "Edu_mother" represent the education qualifications of the student's father and mother. "EcoStatus_family" represents the economic status of the student's family. Regarding to reading related factors, "Readbook" means the reading interest of a student was referred to as Readbook. "Topic" represents the favorite type of book that were coded mathematics/physics as "a", literature as "b", foreign language as "c", natural science/chemistry/biology as "d", history/geography as "e", and information technology as "f". "TimeScience" indicates the length of time students spend reading science books daily, which were coded as follows: less than 30 minutes = 1, between 30 and 60 minutes = 2, and over one hour = 3.

As for the first research question, we use variables from the category of family background such as "EcoStatus_family" and "Career_father(factor)" as independent variables and "Quiz" scores as a dependent variable. In the second research question, we include “Readbook” and “Topic” related to student reading preference as independent variables. “Quiz” scores is a variable that responds to the change. For the last question, we would like to study the relationship between “Quiz”, “TimeScience” and “Career_goal”, thus no independent and dependent variables applied in this case.

2.3 EDM Methods

For the first question, the chosen EDM methods are clustering (using the k-means algorithm) and visualization (using a conditional boxplot with a superimposed violin plot). Clustering guarantees convergence that it can be used to group students with similar characteristics into clusters, which can then be visualized using a boxplot to compare the differences in quiz scores among the different clusters. The violin plot can be used to show the distribution of quiz scores within each cluster, providing additional insight into the relationship between family background and student learning. In this study, we consider clustering and visualization are suitable for the first question because they can help to identify groups of students with similar family backgrounds and patterns in student quiz scores.

For the second question, the chosen EDM methods are predictive analysis (using decision trees) and visualization (using a scatter plot). Decision trees can be used to build a model that predicts educational achievements based on children's reading habits, and the model can then be evaluated to determine its accuracy and reliability. A scatter plot can be used to visualize the relationship between reading habits and educational achievements, providing a visual representation of the model's predictions. From our perspective, Decision Tree and Scatter Plot are appropriate for the second research question because they can be used to build a model that predicts educational achievements based on children's reading habits and evaluate its performance.

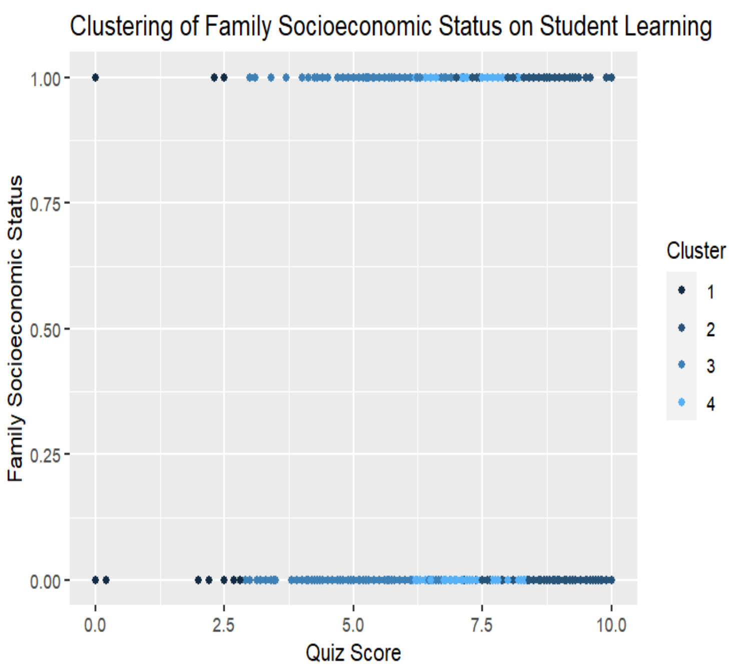
For the third question, the chosen EDM methods are visualization (using a scatter plot), and logistic regression. A scatterplot can be used to visualize the relationship between STEM performance, learning habits, and career choices, providing insight into the factors that influence students' career decisions. Logistic regression can be used to build a model that predicts career choices based on STEM performance and learning habits. Because of the features such as simplicity, flexibility and adaptability of these mining methods, we decided to apply logistic regression in the last research question. It is useful for identifying the factors that influence students’ career choices and building predictive models for career choices.

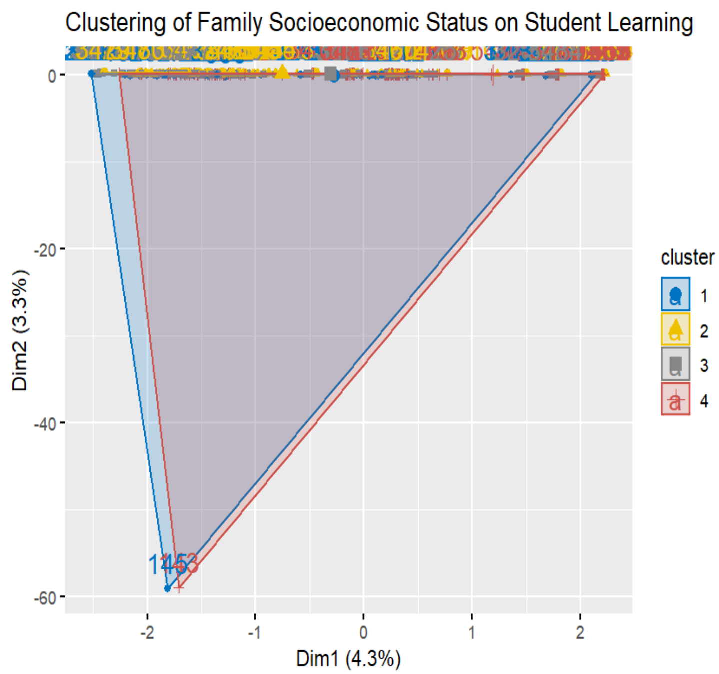
2.4 Data Analysis

As for data analysis methods in the first research question, we apply clustering (kmeans) and visualization.

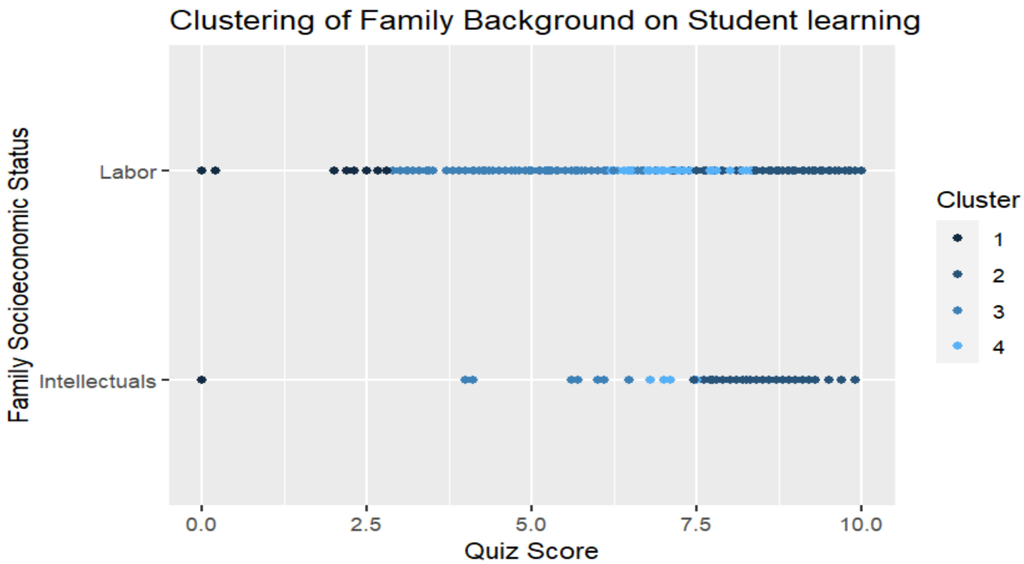
1. *Cluster Analysis (kmeans)*

For modeling the relationship between socio-economic background of students and their performance in class we chose cluster analysis, because we wanted to analyze and divide students on the basis of their performance in class and their economic background. The K nearest neighbors algorithm seems to be a good tool to model students’ academic performance as the clusters divided the students in different groups based on how good their academic standing is and that can be related with their financial background. The only limitation of this method was that it was not adaptive with numerical data, it was beneficial for categorical data.





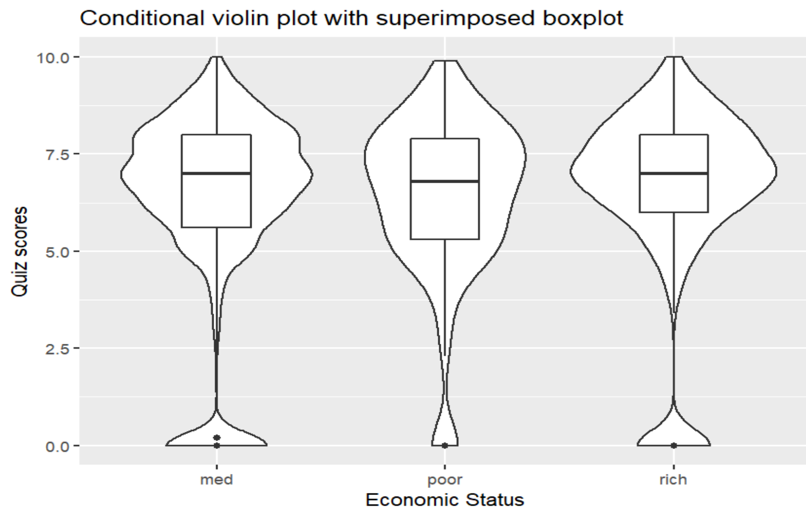
The clusters divided in the above graph are based on the quiz scores of the students, with cluster 1 indicating the students with highest scores and cluster 4 indicating the students with the lowest scores. The socio-economic status is denoted by the economic status of the student's family, this is a dummy variable with 0 denoting poor and 1 denoting rich. The clusters for high and low scores are overlapping in the graphs above, for the first graph there are more high score students in the family with rich economic background but it is not significantly different from the number of the high score students in the family with poor economic background. In the second graph the cluster of high score students has a little overlap with the low score students, but there is an effect of the family income on the scores therefore there is not a complete overlap of clusters.



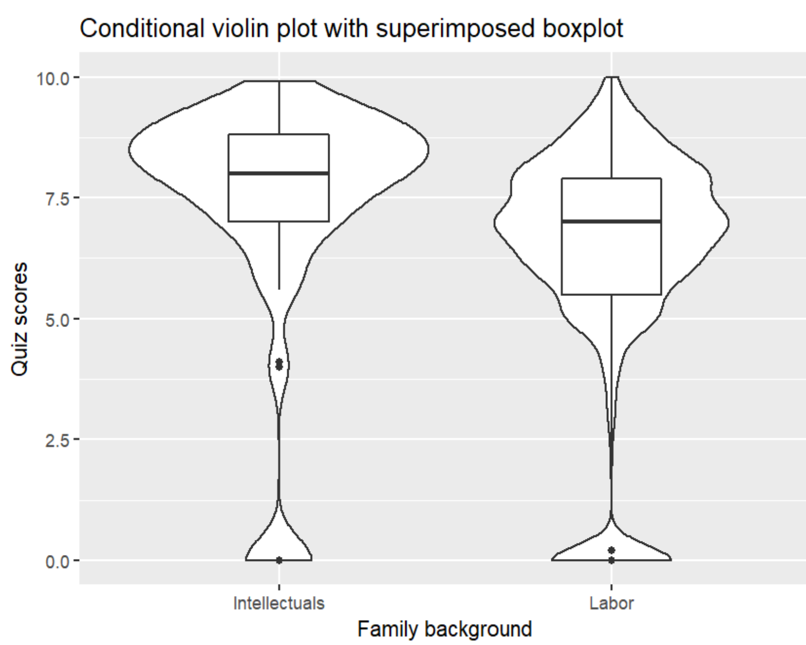
The family background i.e. the profession of the students' parents affect the academic performance of the students. The clusters formed are the academic grades of students, with cluster 1 denoting the highest grades and cluster 4 denoting the lowest grades. The visualization above denotes that the students whose parents are intellectuals have more clusters with high grades and the students whose family background does not have intellectuals have more clusters with lower grades.

2. *Visualization (Conditional violin plot with superimposed boxplot)*

This method of visualization was chosen because boxplot is a very useful tool to plot and compare categorical data vs numerical data. The dataset had factors which needed to be compared to the numerical average of scores therefore this plot was used. The violin plot which is superimposed on the boxplot effectively conveys the distribution of quiz scores. The only limitation that I faced while using this method for visualization was that more variables as factors cannot be accounted for in boxplot, it just uses and compares a single categorical variable with the given dependent numerical variable.



The visualization of conditional boxplot indicates the difference in sample average and density of scores on the basis of the economic status of the students, the average score of the students belonging to rich families is greater than the other two, but there is not a significant difference. Interestingly, students belonging to poor families have the least number of students scoring a minimum score. This method of visualization effectively conveys the distribution based on factors of the data set and has the limitation that the dependent variable can only be categorical.

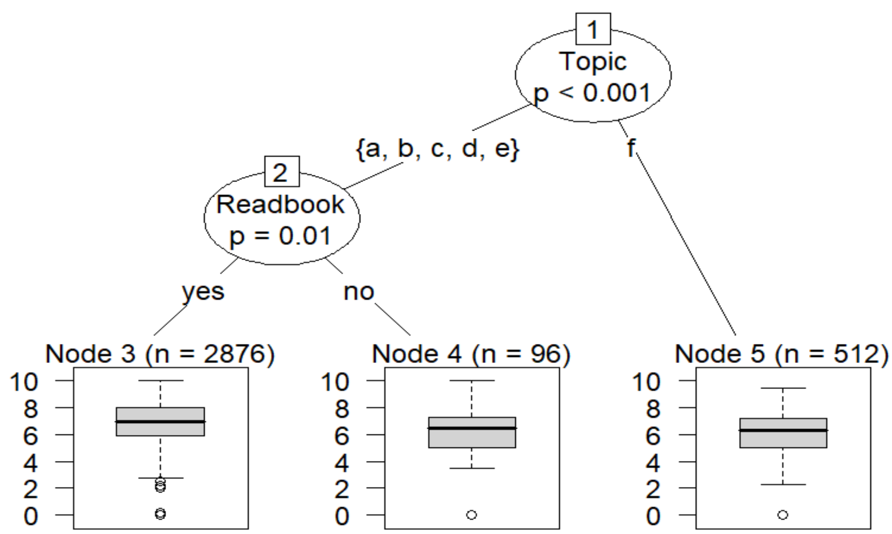


The visualization of conditional plot indicates that the family background of students has a huge impact on the performance of students in class. The students belonging to a family where the parents are intellectuals are more susceptible to perform well in their class, as the average scores obtained by these students is much higher than students with their parents pursuing the labor profession.

As for data analysis methods in the second research question, we apply classification (decision trees) and visualization.

1. *Classification (Decision Trees)*

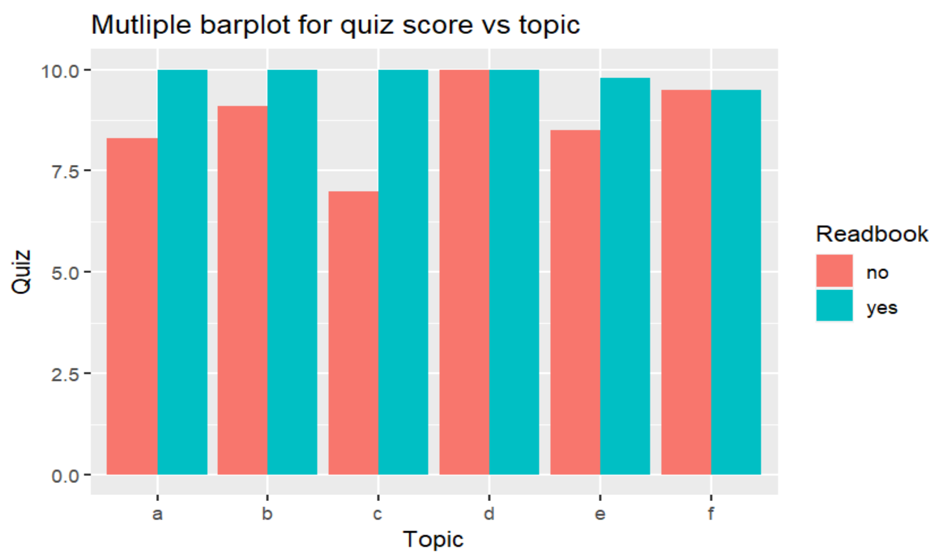
The method of classification using decision trees is a useful method to classify the marks obtained by students based on the factors such as reading preferences and the topics students preferred when they were reading.



The decision tree method actually uses the topic of preference of students and their reading habit to classify and predict the marks obtained by students in the quiz. The topics a,b,c,d and e denote mathematics/physics, literature, foreign language, natural sciences and history/geography respectively, and the readbook variable indicates preference of students for reading a book. From the decision tree obtained above it can be seen that the students favoring topic f which is related to information technology have a lower average score than the students favoring other topics. Furthermore, if the student shows interest in reading books then it can be clearly observed that on an average if students are reading books then they are obtaining better grades than the students who are not reading books.

2. *Visualization*

The visualization technique chosen is multiple barplot which captures and effectively visualizes the quiz scores of students based on preferred topic of study and their preference of reading books.



The visualization clearly showcases the quiz scores of the students based on their choice of reading books for various topics. For topics a,b,c,d,e and f the scores obtained by students when they chose to read books were higher than the score of students who did not read books. Therefore reading books significantly improved the scores of students.

As for data analysis methods in the third research question, we apply logistic regression and visualization.

1. *Logistic regression analysis*

Logistic regression analysis is used to measure linear relationship between two variables, since the response variable is a binary in this case, logistic regression seems like a good fit for analyzing their relationship.

	stem_factorbin		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.53	1.27 – 1.85	**<0.001**
TimeScience	0.84	0.75 – 0.94	**0.002**
Observations	3484		
R^2 Tjur	0.003		

The coefficient implies a positive relationship between predictor and response variable. The association between predictor and response is statistically significant, where one unit increase in time devoted to learning STEM subjects increases the log odds for opting for stem career to 0.84 units. This means that if the students' learning time is devoted more towards STEM fields then there is a higher chance that the students will pursue a career in STEM.

	stem_factorbin		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.47	1.23 – 1.77	**<0.001**
Quiz	0.96	0.94 – 0.99	**0.005**
Observations	3484		
R^2 Tjur	0.002		

The coefficient implies a positive relationship between predictor and response variable. The association between predictor and response is statistically significant, where one unit increase in quiz score of students increases the log odds for opting for stem career to 0.96 units. This means that if the student performs well academically, then the chance that the students will pursue a career in STEM increases. Therefore it can be concluded from the regression analysis that the students who perform well academically in evaluations and devote time towards the study of STEM subjects tend towards pursuing a career in STEM.

2. *Visualization*

The visualization depicts the relation between career choices made by students based on their academic performance and time devoted to studying STEM subjects. The binary variable 0 denotes STEM field and 1 denotes Non-STEM field, the time devoted for studying maximum for label 3 and minimum for 1. Therefore, it can be inferred that the students who want to opt for the STEM field have their quiz score clustered more towards the higher values than the students who want to pursue non STEM fields. The students gaining the highest marks are also the ones who are putting maximum hours studying for the subject with some exceptions.

3. RESULTS OR FINDINGS

3.1 Results of Research Question #1 To what extent does a family's socioeconomic background influence student learning?

The socioeconomic factors such as family's economic status and background were taken into account to determine if they have an effect on the academic performance of students. The economic factor did not have a significant impact on the students' performance, but the parents' profession played a significant role in improving the students' academic performance. The profession of parents play a major role in impacting students to achieve academic success, as the parents who are intellectuals guide their children and motivate them for achieving and doing well in their academic endeavors.

3.2 Results of Research Question #2 Do student's reading habits influence their educational achievements?

The students' learning and reading habits such as reading books based on their favorite topics was indicative of a good academic performance. Regardless of the topic of interest in the books students were reading, if the students inculcate the habit of reading then it will benefit them in attaining a better academic score. Therefore, this learning habit was beneficial for students for improving their educational achievements.

3.3 Results of Research Question #3 What is the relationship between students' STEM performance, learning habits and career choices?

The factor of devoting time to studying STEM courses and future career prospect of pursuing a career in STEM were associated. The students who made a career choice of pursuing STEM were the students who were in a good academic standing and made conscious efforts to devote their maximum time in learning STEM subjects. Therefore in order to promote STEM careers among students, they need to be encouraged to put more effort in improving academic performance through putting more time and effort in learning STEM subjects.

DISCUSSIONS AND CONCLUSIONS

This study aims to study factors influencing student learning and career decisions in the STEM field. We study the influence of family background on student academic performance, the impact of reading habits on student academic performance, and the relationship between student test scores, time spent on reading science books daily, and future career goals.

The results from the k-means algorithm in the first study indicate that parents' occupation influences students' test scores. For instance, students whose parents are intellectuals tend to have children with high grades, and parents whose occupation related to labors are more likely to be associated with children with lower test scores. As for the second study, we applied the decision tree as the data mining method, of which results show students who show the habit of reading books tend to obtain better grades than those who do not read. We conclude that students' reading habits have an impact on their academic performance, while the preference on STEM books has no impact on their test scores. When it comes to the last study, we implemented logistic regression to find out students who devoted more time to science books and perform well in the test tend to show preference on pursuing careers related to the STEM field.

Our scope of study is limited based on the datasets that we have drawn from Vuong (2021). This study is less likely to be referred to beyond the context that the results we obtained can not be generalized for the learning outcomes and career decisions of secondary school students outside of Vietnam. Given the applicability of the data, the issue exists that we need to be aware of in the future research. The quiz scores are limited for single performance regarding students' learning outcomes in the STEM subjects. Therefore, it would be not conclusive to generalize for student overall academic performance. We call for a need for more research studies on student learning outcomes using their learning behavior and effect of extracurricular activities on academic performance.

APPENDIX

The programming language that was used for visualization, data analysis and data transformation is R. Therefore, following are the R codes which were used to obtain the above EDM models and visualizations.

```
# Research Question 1
# The relationship between family's socioeconomic background and student learning.
# Convert into dummy variables
data_dummies <- model.matrix(~ Quiz + Career_father + Career_mother + EcoStatus_family,
Final_data)
view(data_dummies)
#Clustering
kmeans_result <- kmeans(data_dummies, 4)
#Print cluster results
print(kmeans_result$cluster)
#Plot the clusters
plot(Final_data, col= kmeans_result$cluster)

library(ggplot2)

# Check for constant or zero columns
constant_columns <- which(apply(data_dummies, 2, function(x) length(unique(x)) == 1))
# Remove constant or zero columns
data_dummies <- data_dummies[, -constant_columns]
# Convert data_dummies to data frame
data_dummies_df <- as.data.frame(data_dummies)
data_dummies_df <- data_dummies_df %>%
  mutate(EcoStatus_family = EcoStatus_familypoor + EcoStatus_familyrich)

# Visualize the data with ggplot2
ggplot(data = data_dummies_df, aes(x = Quiz, y = EcoStatus_family, color = kmeans_result$cluster)) +
  geom_point() +
  labs(title = "Clustering of Family Socioeconomic Status on Student Learning") +
  xlab("Quiz Score") +
  ylab("Family Socioeconomic Status") +
  guides(color = guide_legend(title = "Cluster"))

# Visualize the results of the clustering analysis using fviz_cluster()
fviz_cluster(kmeans_result, data = data_dummies,
        palette = "jco",
        ellipse.type = "convex",
        labelsize = 15,
        main = "Clustering of Family Socioeconomic Status on Student Learning")

# Visualize the data with ggplot2
```

```
ggplot(data = Final_data, aes(x = Quiz, y = Career_father.factor., color = kmeans_result$cluster)) +
  geom_point() +
  labs(title = "Clustering of Family Background on Student learning") +
  xlab("Quiz Score") +
  ylab("Family Socioeconomic Status") +
  guides(color = guide_legend(title = "Cluster"))

# Visualize the results of the clustering analysis using fviz_cluster
fviz_cluster(kmeans_result, data = data_dummies,
             palette = "jco",
             ellipse.type = "convex",
             labelsize = 15,
             main = "Clustering of Family Background on Student Learning")

#Visualization of conditional boxplot for economic status vs grades
ggplot(data=Final_data,mapping=aes(x=EcoStatus_family,
y=Quiz))+geom_violin()+geom_boxplot(width=0.3)+xlab("Economic Status")+ylab("Quiz
scores")+ggtitle("Conditional violin plot with superimposed boxplot")
ggplot(data=Final_data,mapping=aes(x=Career_father.factor.,
y=Quiz))+geom_violin()+geom_boxplot(width=0.3)+xlab("Family background")+ylab("Quiz
scores")+ggtitle("Conditional violin plot with superimposed boxplot")

library(stats)
library(caret)
career <- cbind(Final_data$Hobby, Final_data$Career_father.factor.)

Final_data$Career_father.factor. <- as.factor(Final_data$Career_father.factor.)
model <- kmeans(career, centers = 2)

#Research Question 2
# What is the relationship between children's reading habits and educational achievements?
library(party)
library(rpart.plot)

#Convert to factor
Final_data$Readbook <- as.factor(Final_data$Readbook)
Final_data$Topic <- as.factor(Final_data$Topic)
Final_data$EncourAct <- as.factor(Final_data$EncourAct)
# Decision Tree Model
model <- ctree(formula= Quiz ~ Readbook + EncourAct + Topic, data= Final_data)
# Plot the decision tree
plot(model)

#Visualization for question 2
library(ggplot2)
ggplot(Final_data, aes(fill=Readbook, y=Quiz, x=Topic)) +
```

```
  geom_bar(position="dodge", stat="identity") +ggtitle("Mutliple barplot for quiz vs topic")

#Research Question 3
library(sjPlot)
library(sjmisc)
library(sjlabelled)

#Logistic regression for question 3
library(dplyr)
Final_data <- Final_data %>% mutate(stem_factor=recode(Career_goal,Doctor="S",
Teacher="S",Engineer="S", Businessman="S", Scientist="S", Lawyer="S", Pilot="S", Accountant="S",
Programmer="S", Architect="S", Astronaut="S", Inventor="S", Hacker="S",
Mathematician="S",Chemist="S",
'Police officer'="NS", Worker="NS", Soldier="NS", Singer="NS", Chef="NS",Banker="NS",
Footballer="NS", Painter="NS", Driver="NS", Actor="NS", 'Tour guide'="NS", 'Fashion
designer'="NS", Interpreter="NS", Writer="NS", Director="NS", Employee="NS", 'Flight
attendant'="NS", Farmer="NS", Athlete="NS", Barber="NS",Repairman="NS", Nurse="NS",
Reporter="NS", Designer="NS", Journalist="NS", Artist="NS", Freelancer="NS", Seller="NS",
Politician="NS", 'Makeup artist'="NS", 'Manga artist'="NS", Gamer="NS", 'Public figure'="NS",
Trainer="NS", Manager="NS", Songwriter="NS", Vet="NS", Firefighter="NS", Model="NS",
Detective="NS", Pharmacist="NS", 'Public servant'="NS", Explorer="NS", 'Martial arts
instructor'="NS", Comedian="NS", Tailor="NS", Secretary="NS", Photographer="NS", Dancer="NS",
Poet="NS", Guard="NS", Lecturer="NS", Swimmer="NS", DJ="NS", Herder="NS", 'Tatoo
artist'="NS", 'Volleyball player'="NS", Diplomat="NS", Legislator="NS",barber="NS",'building
contractors'="NS", 'Public speaker'="NS", Inspector="NS", Choreographer="NS", Philanthropist="NS",
Transporter="NS", Editor="NS", MC="NS",Shaman ="NS",'Chess player'="NS", 'Interior
designer'="NS", 'Toy maker'="NS", Conservationist="NS", Translator="NS",  'Building
contractors'="NS", 'Film director'="NS", 'Stone mining'="NS", Mucisian="NS", Commenter="NS",
Baker="NS", 'Karate instructor'="NS"))

Final_data$stem_factorbin <- factor(Final_data$stem_factor,levels=c("S","NS"),labels=c(0,1))
model <- glm(stem_factorbin~TimeScience, data= Final_data,family="binomial")
summary(model)
tab_model(model)

Final_data$stem_factorbin <- factor(Final_data$stem_factor,levels=c("S","NS"),labels=c(0,1))
model <- glm(stem_factorbin~Quiz, data= Final_data,family="binomial")
summary(model)
tab_model(model)


#Visualization for question  3
library(ggplot2)
Final_data$Time_factor <- factor(Final_data$TimeScience,levels=c(1,2,3),labels=c("1","2","3"))
ggplot(Final_data, aes(x = Quiz, y = stem_factorbin, color=Time_factor)) +
  geom_point(size = 2) +
```

```
  geom_abline(intercept = model$coefficients[1], slope = model$coefficients[2], color = "red", size =
0.5) +
  labs(title = "Quiz Scores vs. Future prospect with respect to time spent",
       x = "Quiz score",
       y = "Future career prospect")
```

REFERENCES

Braun, H. I., White, L. E., & Smith, J. K. (2005). The relationship between reading and science achievement. Reading and Writing, 18(6), 647-671.

Cascio, E. U., & Schanzenbach, D. W. (2015). Family income, child development, and academic achievement. Handbook of Child Psychology and Developmental Science, 7, 1-24.

Charfuelan, M., Wachs, M. A., & Martinez, M. G. (2018). The impact of learning habits on student performance in a learning analytics course. British Journal of Educational Technology, 49(3), 485-496.

Chen, Z., Zhang, T., & Zhang, Y. (2018). Exploring the relationship between STEM performance and future job opportunities: An analysis of PISA data. Journal of Science Education and Technology, 27(3), 391-400. https://doi.org/10.1007/s10956-017-9727-8

Duncan, G. J., & Murnane, R. J. (2011). Whither opportunity? Rising inequality, schools, and children's life chances. New York, NY: Russell Sage Foundation.

Durmaz, F., Eren, E., & Aydin, K. (2017). Predicting student success in learning analytics: The role of learning habits and self-regulated learning. Turkish Online Journal of Educational Technology, 16(4), 189-197.

Fang, Z., & Wei, X. (2008). The impact of reading and writing on science achievement: A meta-analysis. Journal of Research in Science Teaching, 45(1), 26-44.

Hayes, K. E., Zeytinoglu, S. E. C., & Dalton, M. E. (2012). The role of science identity and science interest in the STEM career choice process. Research in Science Education, 42(1), 49-65.

Nores, M., & Barnett, W. S. (2005). The effects of parental education and family income on child achievement: The indirect role of parental expectations and the home environment. Journal of Educational Psychology, 97(1), 44-53.

Ranjeeth, K., Bhat, K., & Nanjundaiah, R. (2017). The relationship between academic performance and outside-class habits of Indian secondary school students. Data in Brief, 13, 721-726.

Ronfard, S., & Guimond, S. (2018). Long-term effects of family income on children's academic achievement: A meta-analysis. Child Development, 89(2), e133-e152.

Schaller, K. C., & Scott, E. C. (2018). The Relationship Between High School Science and Mathematics Coursetaking and Postsecondary STEM Fields of Study. Journal of Research in Science Teaching, 55(4), 479-497. https://doi.org/10.1002/tea.21451

Shariff, S., Campbell, J. R., & Cote, S. G. (2017). The role of learning habits in predicting student success in learning analytics. The Internet and Higher Education, 33, 22-31.

Taspinar, O., Cakir, G., & Eren, M. (2018). The role of learning habits and self-regulated learning in predicting student success in learning analytics. Turkish Online Journal of Distance Education, 19(4), 63-74.

Vuong, Q., La, V., Ho, T., Pham, T., Vuong, T., Vuong, H., & Nguyen, M. (2021). A Data Collection on Secondary School Students' STEM Performance and Reading Practices in an Emerging Country. *Data Intelligence, (3).* 336-356.

Welsh, J. A., & Lippman, T. M. (2009). The relationship between parental education and student achievement. The Journal of Educational Research, 102(5), 335-345.