# Name: Kohn, Aaron Date: 02/15/2021 Title: Understanding Student Performance **Section 1**

## Introduction

Improving the education of children is a goal all parents and educators share. Understanding what factors are related to success should help identify where resources should be focused. What effect does background and social-economic status have on a student’s success? Which students are more likely to need help? What can be done to increase the chances for success?

## Research questions

* Can a student’s level of success be predicted based on their background?
* Which factor in a student’s background has the strongest effect on performance?
* Are there other factors out of school that are related to success?
* Can one or a group of factors be identified that can improve student’s performance?
* How does a parent’s involvement relate to the student’s performance?
* What characteristics can be associated with successful students?
* Are there habits that can be encouraged to help students with lower predicted results?

## Approach

My approach to trying to analyze the data includes a few steps. First, I would run a correlation analysis to determine which factors are related to success. This step would include determining if any other factors are related. Then an attempt would be made to find a model that can accurately predict results. The focus would be on linear regression models and k-nearest neighbors.

## How your approach addresses (fully or partially) the problem.

Determining relationships between factors and student success is the main task. Finding an accurate model that predicts success is a good way to determine which factors combine to help predict better results. A final analysis can be made to see if students whose background may indicate a lower success rate, have specific habits that can be implemented to improve grades.

## Data

* Datasets,
  1. The Students’ Performance in Exams dataset was retrieved from, <https://www.kaggle.com/spscientist/students-performance-in-exams>.
  2. The Students' Academic Performance Dataset (Amrieh, Hamtini, & Aljarah) was retrieved from, <https://www.kaggle.com/aljarah/xAPI-Edu-Data>.
  3. The Student Performance Data Set (Cortez & Silva) was retrieved from, [Student Performance Data Set | Kaggle](https://www.kaggle.com/larsen0966/student-performance-data-set) and [Math Students | Kaggle](https://www.kaggle.com/janiobachmann/math-students).
* Data details
  1. The first dataset contains 1000 observations of 8 variables. Three of which are student marks in reading, writing and math. There does not appear to be any missing data. Of note is the lunch variable which contains the values standard or free/reduced. This may be of interest since school subsidies are usually based on income levels.
  2. The second dataset consists of 480 instances of 17 variables. The final one being the grade result classified into three levels, low level includes grades 0 – 69, mid-level for grades 70 – 89, and high level for grades 90 – 100. The data was collected using a learner activity tracker tool, that can monitor student activities.
  3. The third dataset consists of two files one of 649 observations of 33 variables for Portuguese language marks. The final three G1, G2, and G3 are grading period marks. First period, second period and the final mark respectively. The information was collected using school records and questioners. The second containing math course information only contains 395 observations.

## Required Packages

car

caTools

class

ggplot2

Hmisc

QuantPsyc

This list is preliminary and will be updated as the project proceeds.

## Plots and Table Needs

Scatter plots with model lines can be used to show relationships. Histograms and boxplots will be used to model distributions and outliers. A heat map can be used to illustrate correlation results.

## Questions for future steps.

While I am somewhat confidant analyzing correlations and have managed to create functional models, I am less comfortable when dealing with non-numeric values. Understanding the results and steps to be taken to improve the model is still a process that I don’t fully understand. One final question is whether it is acceptable to use a k-means algorithm on data with known results in order to determine groupings that can be used to predict results.

# References

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.