**DATS 6101-11 Introduction to Data Science**

**First Project (Fall 2025): G#3**

**Dataset and Research Proposal**

This proposal outlines a research project focused on analyzing the factors that influence student academic performance. The study will utilize a synthetic dataset (obtained from [Kaggle](https://www.kaggle.com)) containing 1,000,000 rows and 6 columns with the variables: **student\_id, weekly\_self\_study\_hours, attendance\_percentage, class\_participation, total\_score**, and **grade**. With all numerics, except the header row and grade column, makes the data set ideal for exploratory data analysis and deriving insights to understand the performance of each student. By quantifying the data and applying relevant statistical tools such as summary statistics, descriptive statistics, ggplot2 among others, we can better perceive the data and make sound conclusions. There is research that considers other factors of student performance such as nutrition and student performance by Taras Howard (2005); indoor air quality and student performance by the United States Environmental Protection Agency (2003); education resources and student performance by Yaniv, R and Adi, Y (2021). They are all mostly external factors that can affect the performance of a student. Our study narrows the focus down to the internal factors or what could be described as accessible by all students against their total scores.

**Dataset Source:** The dataset can be found on Kaggle at the following link

[Student Performance Dataset](https://www.kaggle.com/datasets/nabeelqureshitiii/student-performance-dataset?resource=download)

**Research Topic**: **Analyse the Factors that Contribute to Student Performance**

**SMART Questions of the Research**

1. Do students with **attendance ≥ 85%** achieve higher average **total scores** than those with attendance < 85%?
2. Is there a **linear relationship** between the number of weekly self-study hours and a student’s total\_score?
3. Are there any students who dedicate very few hours to studying but still achieve a higher  total\_score?
4. What are the average weekly study hours, attendance percentage,class  participation and total scores for all students in the dataset?
5. Do students with unusually high total\_score despite low study hours, attendance, and participation show significant anomalies?

**Reference**

Taras, H. (2005), Nutrition and Student Performance at School. Journal of School Health, 75: 199-213. <https://doi.org/10.1111/j.1746-1561.2005.tb06674.x>

United States Environmental Protection Agency (2003). Indoor Environmental Division. *Indoor Air Quality & Student Performance.*  Washington, D.C: U.S.

Yaniv, R. and Shany, A. (2021), “Education Resources and Student Performance: Evidence from Arab Municipalities in Israel.” *Applied economics* 53.38): 4386–4403.

Nabeel, G.M (n.d). Student Performance Dataset (Synthetic, Realistic). Kaggle