**Basic regression analysis**

Pre-requisite: We assume that the student has a running python environment with Jupiter notebook.

Introduction:

In this assignment, you will conduct a regression analysis on the Student performance dataset to gain insights into the relationship between various independent variables and the target variable, such as exam scores or academic achievement. The goal is to understand the impact of different predictors on the target variable and assess the performance of linear regression models.

In the assignment, we assume that you have performed the basis exploratory analysis on the subset of student performance, because you are using your assumptions and conclusions on the subset as a starting point to this regression analysis.

Note: You will be using the same subset of student performance you used during basis verkennende analyse assignment.

Task1: Load the data and separate columns

* Read the csv file in a pandas dataframe
* Distinguish numerical and categorical variables: It is a good practice prior to preprocessing store variables in separate dataframes based on their nature. Later you can apply preprocessing technics to the same type of variables.
  + To identify categorical and numerical features, store their names in separate lists.
  + Among categorical features, check which of these can benefit from target encoding and which are nominal and need encoding.

Task 2: Data Preprocessing

At this step, you are preparing the data so that you can use it later for regression. You do this using a pipeline:

* While exploring the subset during the “verkennende analyse opdracht”, you have identified few columns that contain missing values. Let’s start then with creating SimpleImputer objects for each type of data.
* Make custom Transformers to encode your nominal and ordinal categorical features.
* Make individual pipelines for each type of data (continuous, numerical ordinal, categorical nominal, and categorical ordinal) that includes the appropriate imputer and encoder.
* Use the Column Transformer to put the individual pipelines together.
* Make the full pipeline that uses the preprocessor created from ColumnTransformer.

Task 3: Regression Analysis

* As a first step of your regression analysis, you explore the linear regression between the target variable and 1 independent variable.
  + Put in markdowns your interpretation
* Next, you explore linear regression this time between the target variable and other independent variables. For this you need to look at which predictor or set of predictors has the most impact on the target variable.
* Use for your analysis appropriate visualization techniques or statistical tests to assess the strength of the relationship between the target and independent variables.
* Write your assumptions in markdown cells.

Task 4: Identifying the best preprocessing combination for regression

* Identify how you could tune your model to improve performance. For this you can identify which transformations to bring to which columns so that the model performance can improve. Think about using different scaling approaches and encoding approaches for example.
* The best way to identify which preprocessing approach is best for your model performance is by testing different ones with your model using technics like Grid Search. And then decide on the best one based on how the regression model performs.
* Identify the hyperparameters for the preprocessing steps and the values they might take
* Set up a parameter grid for your identified hyperparameters.
* Run your Grid Search with your chosen cross-validation method.
* Once the Grid Search has completed, what combination of parameters was found to give the best results?
* Evaluate your model on the test data.

Task 5: Check other models for regression

* In this part you can use python libraries such as Pycaret to test the performance of multiple models at once.
* Set your interpretations in markdown cells

Note: Ensure that you provide detailed explanations, include necessary comments to your code snippets, and utilize appropriate visualizations to support your analysis at each step.