Name: Waken Cean C. Maclang
Date: September 8, 2025

Problem Scenario:

You are hired as a **junior data scientist** in your university's research office. The Dean wants to know if student study habits (e.g., hours of study, use of online learning platforms, attendance) are related to academic performance.

Tasks:

1. Problem Definition

   Problem: We want to understand how a student's study habits, such are their study hours, attendance, and use of LMS, affect their academic performance?

   Several studies noted how a student's e-learning activity [1] and motivation to learn (e.g., study hours and attendance) [2] significantly affect student performance. Thus, we will use the noted variables as the following:

   Student ID will not be included as it serves as the index of the dataset.

   Dependent Variables: Study_Hours, Attendance, Uses_Moodle

   Independent Variables: Final Grade

2. Data Collection & Description

   Describing each variable:

```
RangeIndex: 30 entries, 0 to 29
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Student_ID   30 non-null     object
 1   Study_Hours  30 non-null     int64
 2   Attendance   30 non-null     int64
 3   Uses_Moodle  30 non-null     object
 4   Final_Grade  30 non-null     int64
dtypes: int64(3), object(2)
memory usage: 1.3+ KB
```

Figure 1: Variables and their types present in the Dataset

   Student_ID (String): ID or Identification of a student

Study_Hours (int): Number of hours a student has studied

Attendance (int): Number of times a student has attended class

Uses_Moodle (boolean): Identifies if the student in the row uses Moodle (true) or not (false)

Final_Grade (int): Final grade of the student.

3. Data Cleaning

As seen once the dataset was loaded, see Figure 1, how columns such as Uses_Moodle was labeled with the incorrect datatype.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Student_ID   30 non-null     object
 1   Study_Hours  30 non-null     int64
 2   Attendance   30 non-null     int64
 3   Uses_Moodle  30 non-null     bool
 4   Final_Grade  30 non-null     int64
dtypes: bool(1), int64(3), object(1)
memory usage: 1.1+ KB
```

Figure 2: After changing the datatypes for Student_ID (str) and Uses_Moodle (bool)

Similarly, we will run “df.isna().sum()” to see if there are null values present in each column.

```
None
Student_ID     0
Study_Hours    0
Attendance     0
Uses_Moodle    0
Final_Grade    0
dtype: int64
```

Figure 3: Counting the number of null values (NA) in the dataset

Lastly, we will check for outliers in the data by utilizing a boxplot as seen in Figures 4a to 4c, there are no outliers found for each column.

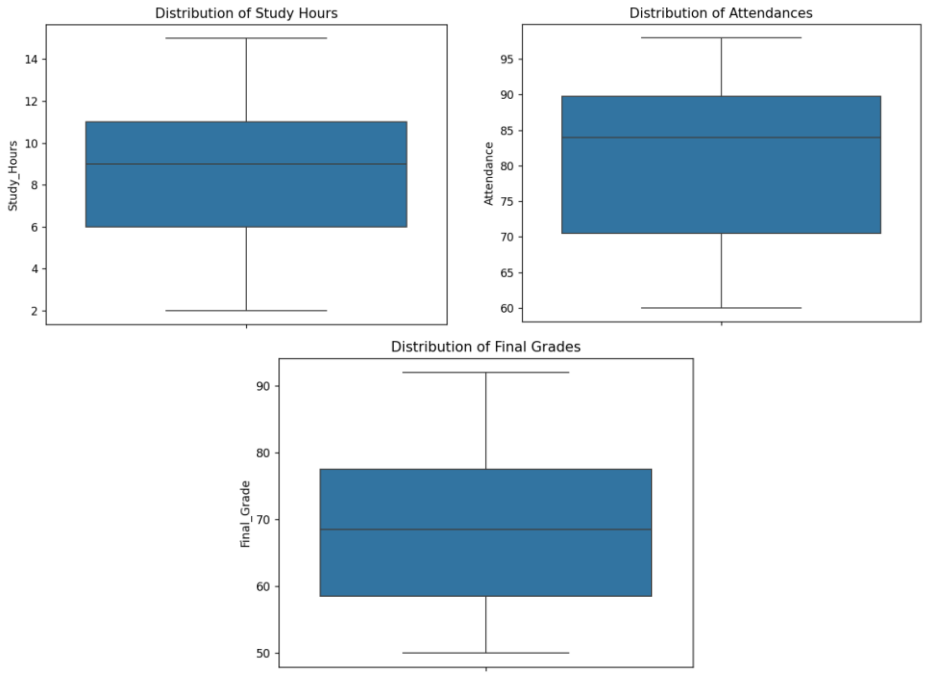

Figure 4: (a) Distribution of Study Hours, (b) Distribution of Attendances, (c) Distribution of Final Grades

4. Exploratory Data Analysis (EDA)

	Study_Hours	Attendance	Final_Grade
count	30.000000	30.000000	30.000000
mean	8.800000	80.733333	68.666667
std	3.497783	11.551782	12.518492
min	2.000000	60.000000	50.000000
25%	6.000000	70.500000	58.500000
50%	9.000000	84.000000	68.500000
75%	11.000000	89.750000	77.500000
max	15.000000	98.000000	92.000000

Figure 5: Measures of Central Tendency and Spread of the Data

The table in Figure 5 shows how there is a large spread among the Final Grades, indicating that there may be a strong influence among study hours and attendance towards a student's final grade, as based on the data collected. Similarly, we check for the correlation among the dependent variables towards our independent variable as seen in Figure 6.

```
Correlation among Study Hours and Final Grade:
[[1.         0.81192487]
 [0.81192487 1.        ]]

Correlation among Attendance and Final Grade:
[[1.         0.87662906]
 [0.87662906 1.        ]]
```

Figure 6: Correlation of Study Hours and Attendance to Final Grade

Figure 6 shows how the two independent variables have a high positive correlation of 0.81 and 0.87 respectively towards the Final Grade, this means, statistically, an increase in study hours or attendance will lead to an increase in a student's final grade, with Attendance having a stronger correlation with Final Grade. This can be clearly viewed in Figures 7a and 7b. We can see how the points are closer to the regression line as compared to the study hours, further solidifying how a student's attendance has a stronger correlation.

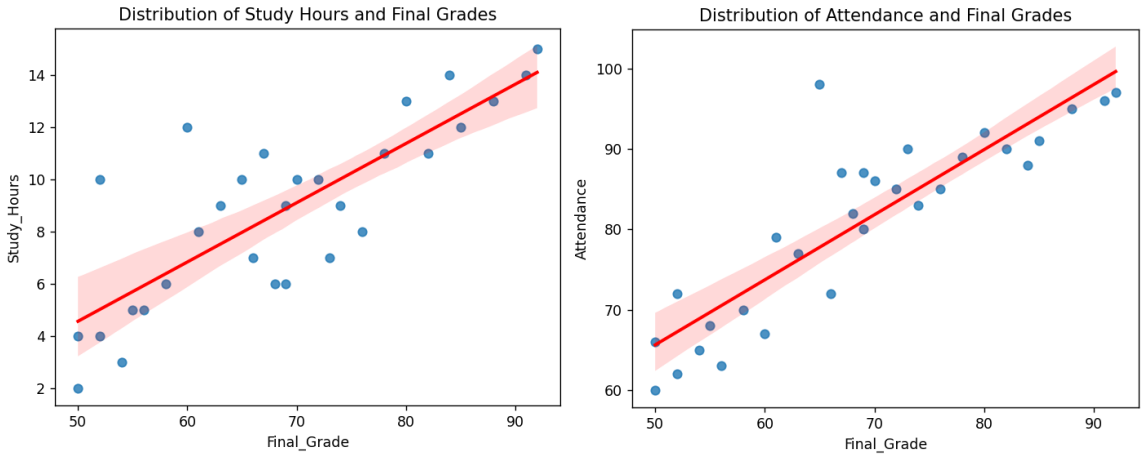

Figure 7: (a) Scatterplot of study hours and Final Grades, (b) Scatterplot of Attendance and Final Grades

Similarly, Figures 8a and 8b further displays the distribution among the attendance and study hours among students.

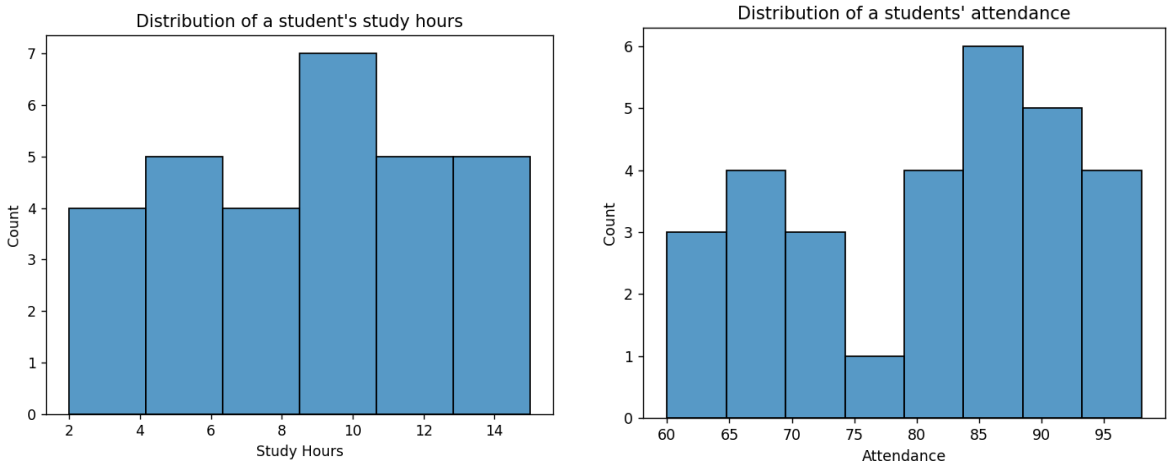

Figure 8: Histogram of students' (a) Study hours (binwidth = 2) and (b) Attendance (binwidth = 5)

5. Modeling

After training the model to predict Final Grade from Study_Hours, the model had an intercept of 28.61 and a regressor of 4.14. This means that the model will predict Final Grade as 28.61 if our regressor (Study_Hours) is 0, and will increase by 4.14 for every point in Study_Hours.

6. Evaluation

After running the evaluation metrics, the simple linear regression model had an R-square of 0.417, meaning that the model given the data, Study_Hours can predict the 41.7% percent of the final grades.

However, this is too low of an accuracy, one way to increase the R-squared score is by adding more features to the data or increasing the records used for linear regression.

Lastly, Linear Regression is limited to the minimum and maximum values of continuous/discrete variables found in the dataset.

7. Deployment – NA

8. Communication

The dataset contained 30 student records consisting of 5 columns, Student_ID, Study_Hours, Attendance, Uses_Moodle, and Final_Grade. From this dataset we want to predict the Final Grade given the data available. We start by first understanding the correlation among independent variables Uses_Moodle, Study_Hours, and Attendance with Final Grade and found that every IV is positively correlated to a student's Final

Grade. This means, statistically, that if a student studies more, attends classes, and uses moodle, the higher the student's grade. However, this may not directly cause a student to have high grades.

We then proceed to creating the linear model, utilizing the Study_Hours column as our regressor and the Final Grade as the value to be predicted, we resulted to having a Linear Regression model that can predict the Final Grades of students 41.7% of the time. Moving on to our regressor and intercept, our model had an intercept of 28.61, which means that if a student doesn't study at all, they would have a grade of 28.61. Additionally, our regressor coefficient is 4.14, which means that for every hour a student studies, their grade increases by 4.14 points.

Our model is limited to the range of our dataset, thus if we try to predict the grade of a student with a Study Hour less than 2 or greater than 15, it may produce unrealistic Grades with others going passed 100. Additionally, the model had a score of 0.417, which is low for a linear regression model.

We recommend incorporating other features in the dataset, such as Attendance, Uses_Moodle, quiz scores, and other predictors of academic performance in the data. Moreover, the model would better predict the Grades of students with more data, as 30 student records is considered as insufficient in this case. Similarly, we could feature engineering in the grades of students to determine which students pass or fail, and use them for a classification model, one that can predict students who are at risk of failing.

Thus, in succeeding models, we would incorporate Multiple Linear Regression (MLR) that can handle multiple predictors to better predict a student's final grades.

References:

[1] Y. Al Husaini, & N. S. A. Shukor (2022). “Factors affecting students’ academic performance: A review”. *Social Science Journal*, vol *12*, no 6, p. 284-294.

[2] J.-T. Seo, B.-N. Bok, and K.-W. Yeon, “Analysis of LMS data of distance Lifelong Learning Center learners and drop-out prediction,” Journal of Human-centric Science and Technology Innovation, vol. 1, no. 3, pp. 23–32, Jul. 2021, doi: 10.21742/jhsti.2021.1.3.04.