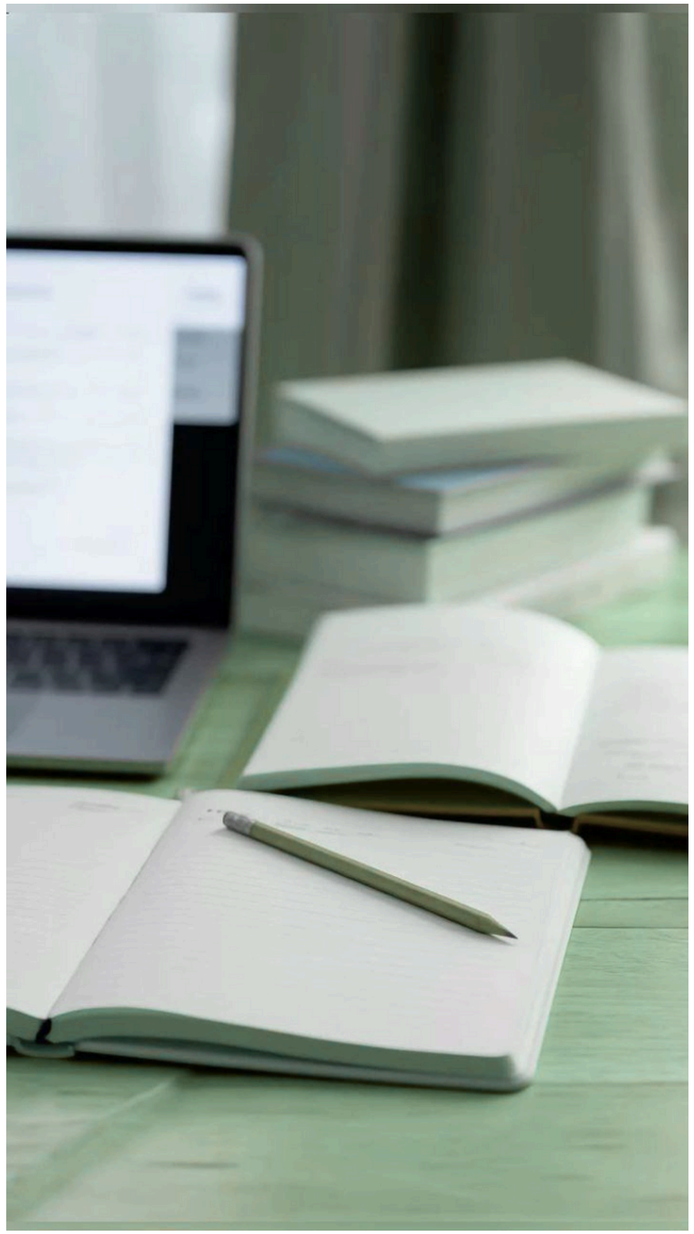


# Students Performance Analysis: Exploring Key Influencing Factors

By: Heba Adel Ali

# Project Overview

- Objective: This project explores how various factors—such as study habits, attendance, past academic records, and demographics—affect student performance.

- Goal: Analyze the data and identify clear patterns and relationships to help build better strategies for educational institutions.

- Methodology: The data was analyzed using Python and data analysis libraries such as Pandas, Seaborn, and Matplotlib.

  This allowed us to:
  Perform Exploratory Data Analysis (EDA) to identify influential factors.
  Construct a machine learning model to predict student exam performance based on these factors.

# Dataset Description

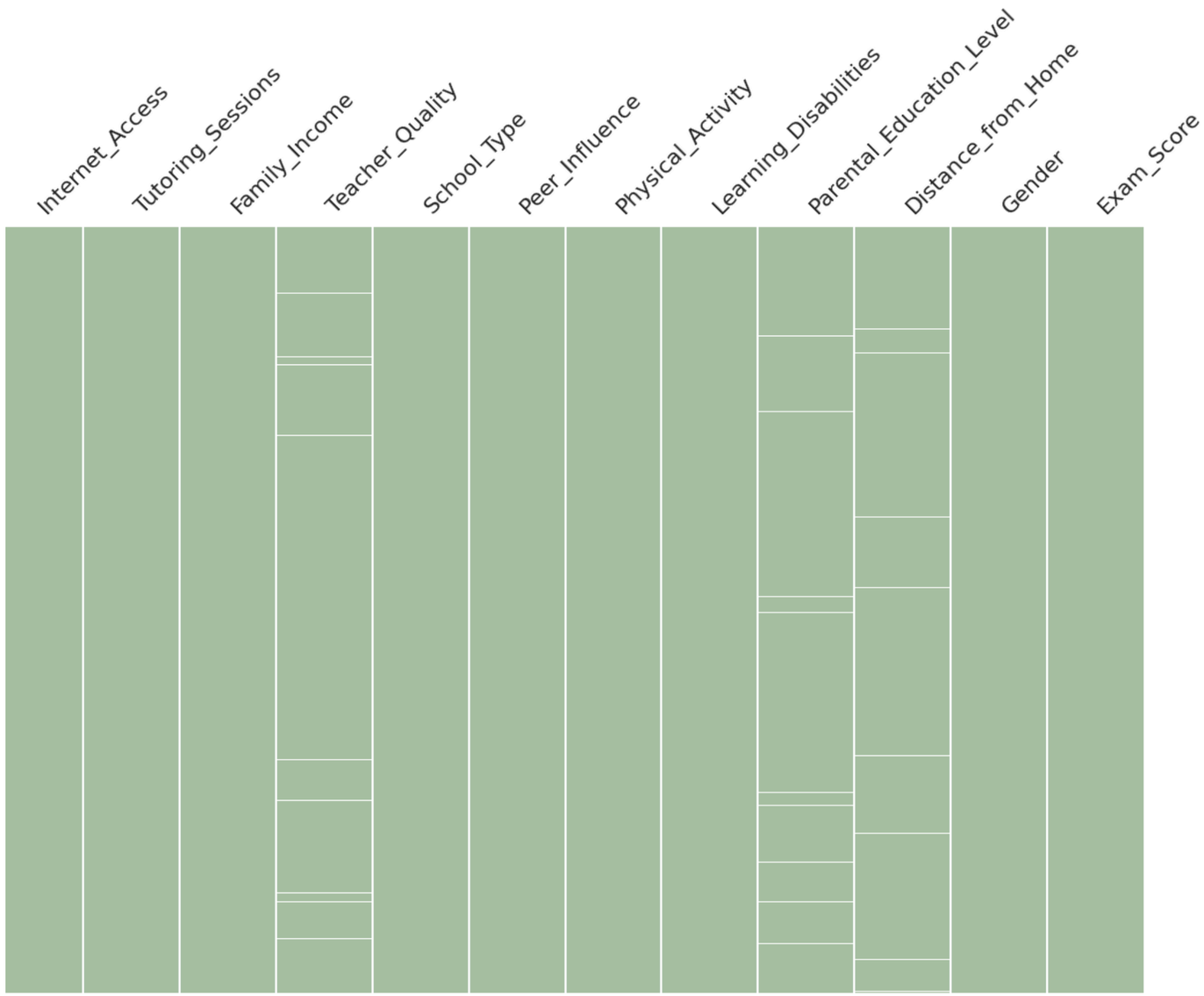
- Shape: 6,607 rows × 20 columns
- Total features: 19 independent variables
- Target: Exam_Score
- Data type mix: Numerical + Categorical

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6607 entries, 0 to 6606
Data columns (total 20 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Hours_Studied               6607 non-null   int64
 1   Attendance                  6607 non-null   int64
 2   Parental_Involvement        6607 non-null   object
 3   Access_to_Resources         6607 non-null   object
 4   Extracurricular_Activities  6607 non-null   object
 5   Sleep_Hours                 6607 non-null   int64
 6   Previous_Scores             6607 non-null   int64
 7   Motivation_Level            6607 non-null   object
 8   Internet_Access             6607 non-null   object
 9   Tutoring_Sessions           6607 non-null   int64
 10  Family_Income               6607 non-null   object
 11  Teacher_Quality             6529 non-null   object
 12  School_Type                 6607 non-null   object
 13  Peer_Influence              6607 non-null   object
 14  Physical_Activity           6607 non-null   int64
 15  Learning_Disabilities       6607 non-null   object
 16  Parental_Education_Level    6517 non-null   object
 17  Distance_from_Home          6540 non-null   object
 18  Gender                      6607 non-null   object
 19  Exam_Score                  6607 non-null   int64
dtypes: int64(7), object(13)
memory usage: 1.0+ MB
```

# Exploratory Data Analysis (EDA)

A few variables had missing data:

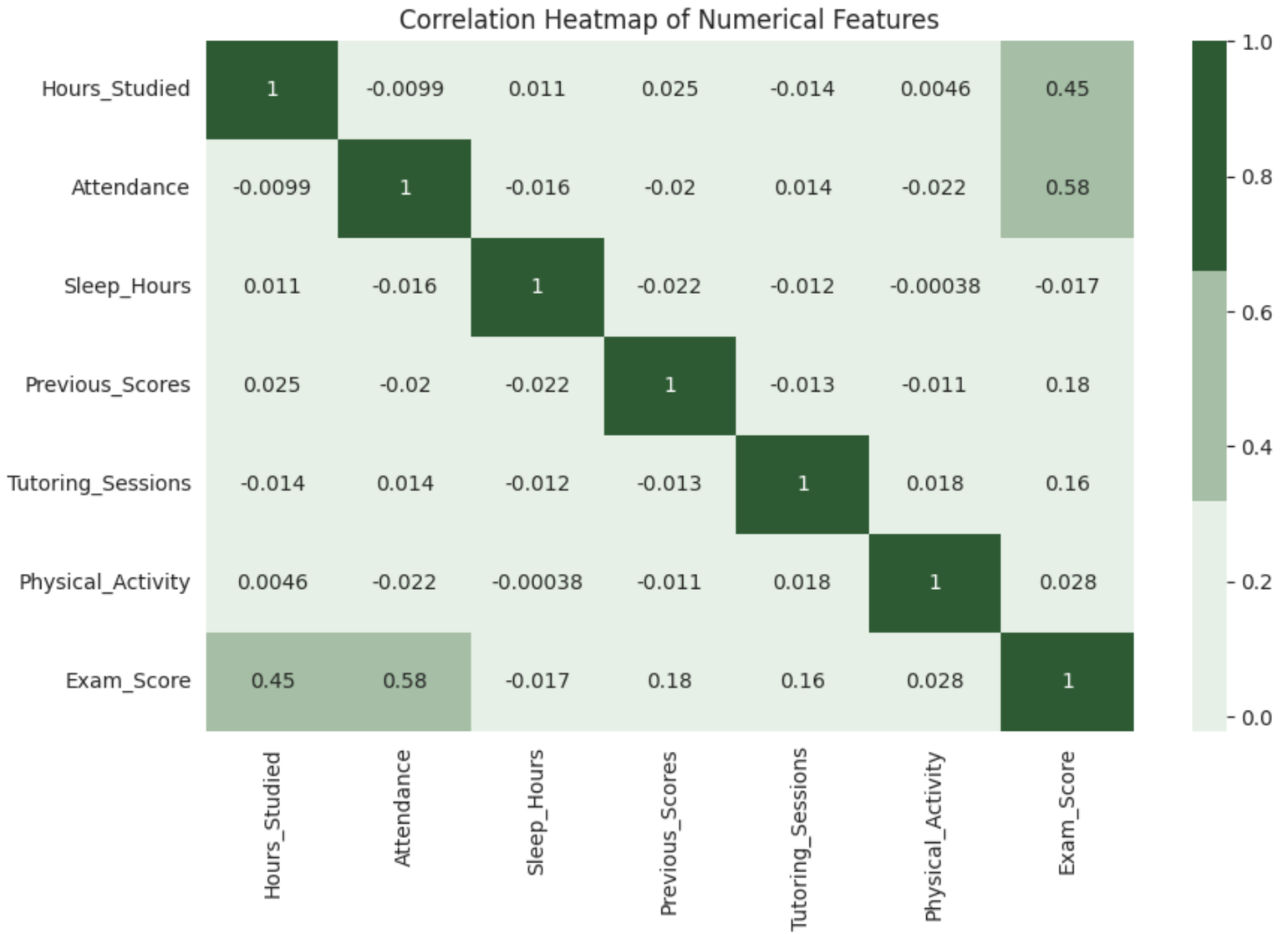
- Teacher_Quality → 78 missing
- Parental_Education_Level → 90 missing
- Distance_from_Home → 67 missing



# Exploratory Data Analysis (EDA)

Insights:

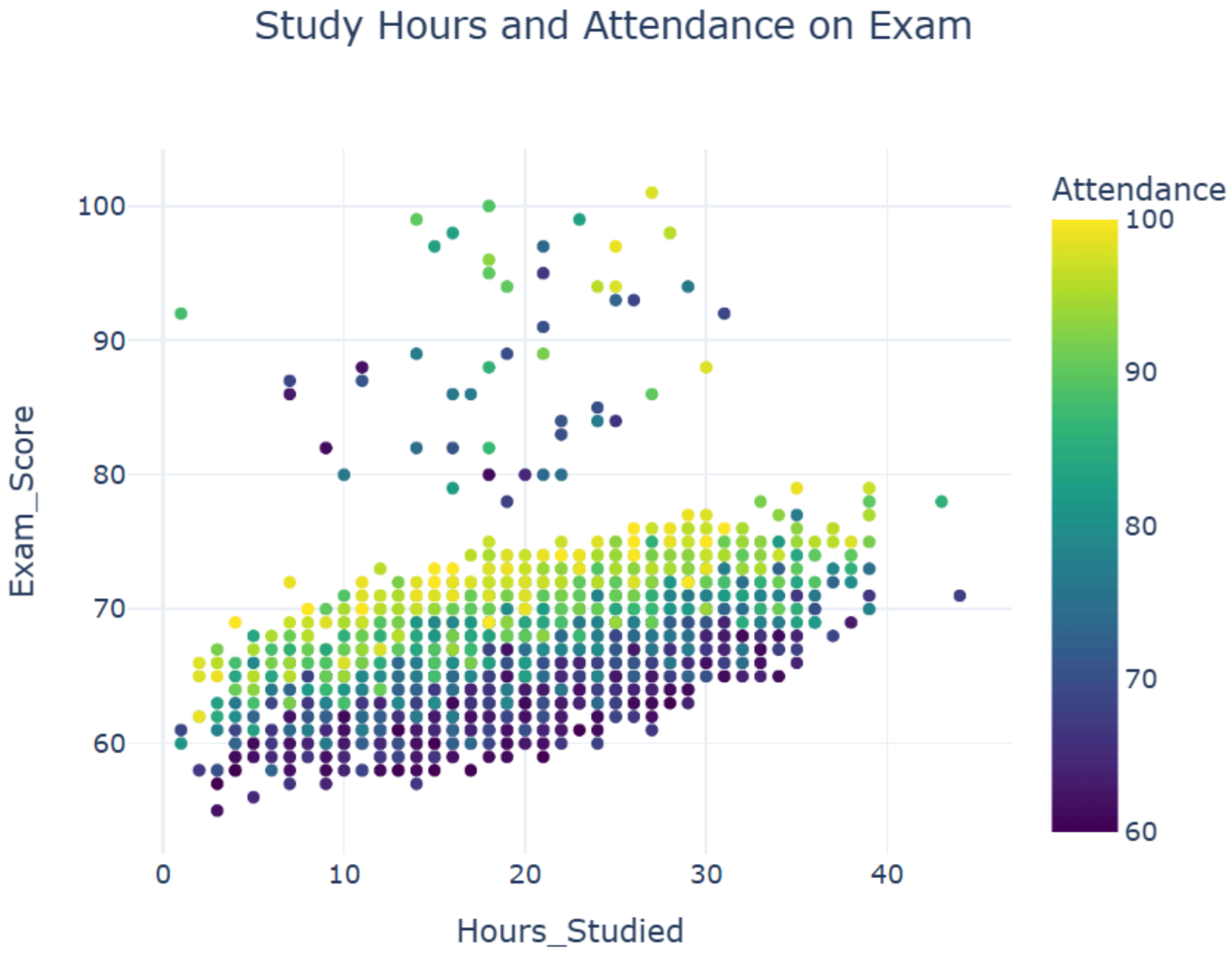
- Exam_Score is most correlated with Attendance (0.58) and Hours_Studied (0.45).
- Other features have weak correlations.



# Exploratory Data Analysis (EDA)

Insights:

- The scatter plot shows a clear positive trend — as Hours Studied increase, Exam Scores also rise.
- Students with higher attendance (yellow dots) also perform better, indicating that both study effort and class participation play a key role in improving performance.



# Data Preprocessing

**Handling Missing Values:** Filled missing entries in Parental_Education_Level, Distance_from_Home, and Teacher_Quality with the most frequent value.

**One-Hot Encoding (OHE):** Converted categorical features into numerical format for modeling.

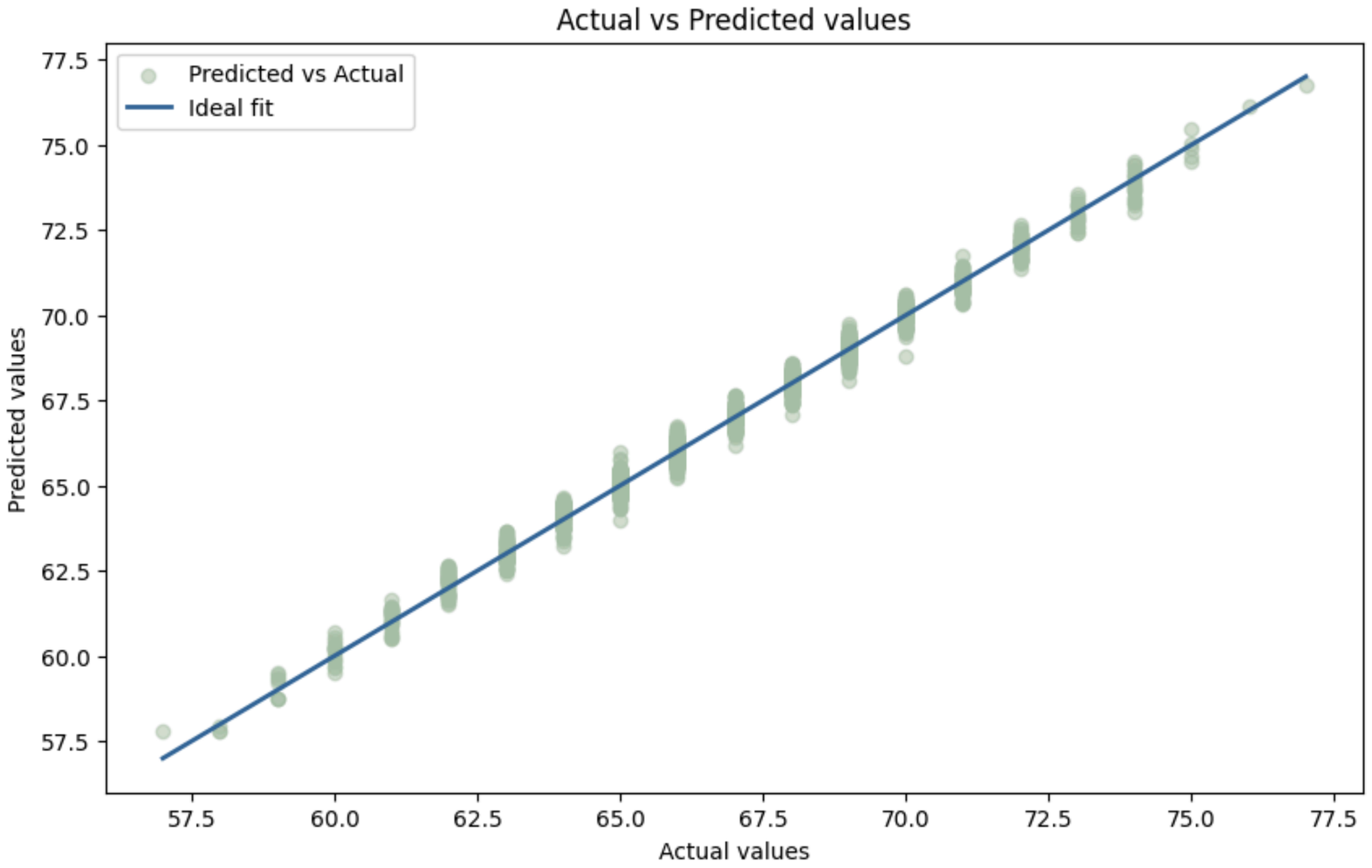
**Outlier Detection (Z-score):** Detected extreme values in numerical columns to assess data quality.

Data is now clean, complete, and ready for model training

# Model Overview

- **Data Split**: Training and testing sets for model evaluation
- **Model**: Linear Regression used to predict Exam_Score.

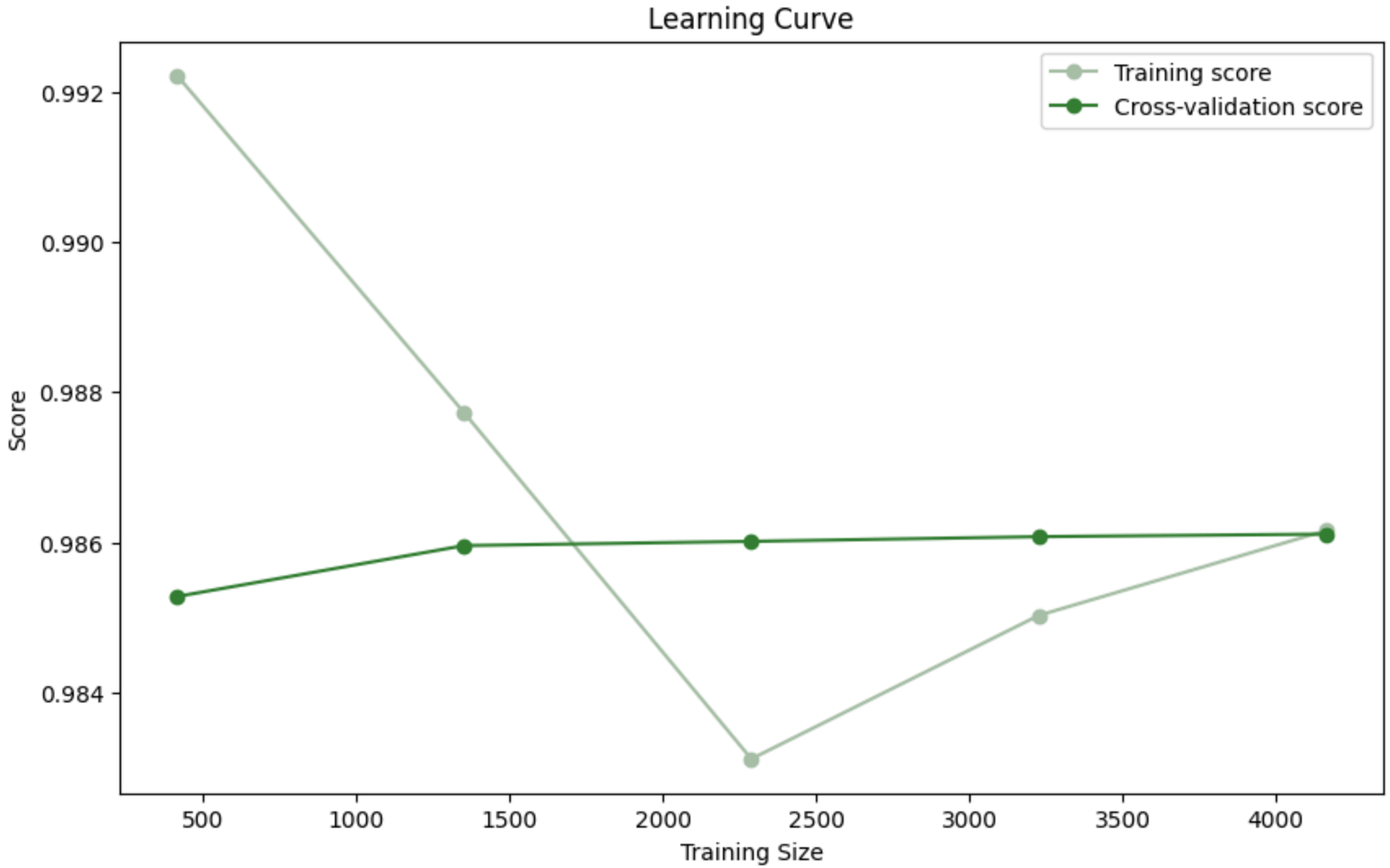
- **Model Evaluation Results**:
  $R^2$ Score (Test Set): 0.99
  Cross-Validation $R^2$: Mean = 0.986
  Error Metrics: MAE = 0.27, RMSE = 0.32



# Model Overview

- **Learning Curve:**

  Shows how training (soft green) and cross-validation (dark green) scores change with training size.
  Indicates the model learns well and generalizes without overfitting.



# Thanks!


Linkedin

hebaadelali

hebadepiii@gmail.com