

# Evaluating Students' Performance Factors

1. Problem Formulation
2. Data Preprocessing
3. Exploratory Data Analysis
4. Modeling
5. Modeling Evaluation & Conclusion

- Goal: Identify key predictors of student academic performance.
- Methods: Data preprocessing, EDA, regression models, clustering.
- Findings: Attendance and hours studied were strongest predictors.
- Impact: Supports early intervention strategies for at-risk students.

# **Given a student's data, what features would best predict their performance and are there patterns that differentiate high-achieving students from the rest?**

Benefits of this topic include:

1. Encourage efficient time management and decrease stress during high-stake academic periods such as final examinations for students.
2. Allow academic staff to develop more effective teaching strategies and apply them to at-risk students for early intervention and support measures (Fishstrom, 2022).

Hours_Studied	Number of hours spent studying per week.
Attendance	Percentage of classes attended.
Parental_Involvement	Level of parental involvement in the student's education (Low, Medium, High).
Access_to_Resources	Availability of educational resources (Low, Medium, High).
Extracurricular_Activities	Participation in extracurricular activities (Yes, No).
Sleep_Hours	Average number of hours of sleep per night.
Previous_Scores	Scores from previous exams.
Motivation_Level	Student's level of motivation (Low, Medium, High).
Internet_Access	Availability of internet access (Yes, No).
Tutoring_Sessions	Number of tutoring sessions attended per month.

Family_Income	Family income level (Low, Medium, High).
Teacher_Quality	Quality of the teachers (Low, Medium, High).
School_Type	Type of school attended (Public, Private).
Peer_Influence	Influence of peers on academic performance (Positive, Neutral, Negative).
Physical_Activity	Average number of hours of physical activity per week.
Learning_Disabilities	Presence of learning disabilities (Yes, No).
Parental_Education_Level	Highest education level of parents (High School, College, Postgraduate).
Distance_from_Home	Distance from home to school (Near, Moderate, Far).
Gender	Gender of the student (Male, Female).
Exam_Score	Final exam score.

# Data Preprocessing

```
#Remove For Duplicate Rows  
data = data.drop_duplicates()
```

```
[ ] #Check For Missing Values  
data.isnull().sum()
```

```
#replace missing values with mean/mode  
data['Distance_from_Home'] = data['Distance_from_Home'].fillna(data['Distance_from_Home'].mode()[0])  
data['Parental_Education_Level'] = data['Parental_Education_Level'].fillna(data['Parental_Education_Level'].mode()[0])  
data['Teacher_Quality'] = data['Teacher_Quality'].fillna(data['Teacher_Quality'].mode()[0])
```

```
#check if exam_score and attendance is bounded [0,100] and remove cells with outliers  
data = data[(data['Exam_Score'] >= 0) & (data['Exam_Score'] <= 100)]  
data = data[(data['Attendance'] >= 0) & (data['Attendance'] <= 100)]
```

To handle preprocessing the data for linear regression:

1. Removed duplicate rows
2. We checked for missing values and replaced the missing values by the mean and modes.
3. Removed rows where the percentage scaled values were not bounded between [0,100]
4. Encoded non-ordinal data by one-hot-encoding and ordinal data by label encoding

```
# Label Encoding For Parental_Involvement
data['Parental_Involvement'] = data['Parental_Involvement'].replace({'None':0, 'Low': 1, 'Medium': 2, 'High': 3})

# Label Encoding For Parental_Education_Level
data['Parental_Education_Level'] = data['Parental_Education_Level'].replace({'None':0, 'Low': 1, 'Medium': 2, 'High': 3})

# Label Encoding For Family_Income_Level
data['Family_Income'] = data['Family_Income'].replace({'Low':0, 'Medium': 1, 'High': 2})

# Label Encoding Parental_Education_Level
data['Parental_Education_Level'] = data['Parental_Education_Level'].replace({'None':0, 'High School': 1, 'College': 2, 'Postgraduate': 3})

# Label Encoding for Access_to_Resources (Low, Medium, High)
data['Access_to_Resources'] = data['Access_to_Resources'].replace({'Low': 1, 'Medium': 2, 'High': 3})

# Label Encoding for Teacher Quality (Low, Medium, High)
data['Access_to_Resources'] = data['Access_to_Resources'].replace({'Low': 1, 'Medium': 2, 'High': 3})

# Label Encoding for Extracurricular_Activities (Yes, No)
data['Teacher_Quality'] = data['Teacher_Quality'].replace({'Low': 1, 'Medium': 2, 'High': 3})

# Label Encoding for Motivationan Level (Low, Medium, High)
data['Motivation_Level'] = data['Motivation_Level'].replace({'Low': 1, 'Medium': 2, 'High': 3})

# Label Encoding for Peer_Influence (Positive, Neutral, Negative)
data['Peer_Influence'] = data['Peer_Influence'].replace({'Negative': 0, 'Neutral': 1, 'Positive': 2})

# Label Encoding for Distance_from_Home (Near, Moderate, Far)
data['Distance_from_Home'] = data['Distance_from_Home'].replace({'Near': 1, 'Moderate': 2, 'Far': 3})

#One Hot Encoding for non-ordinal categorical features
data = pd.get_dummies(data, columns=['Gender', 'Physical_Activity', 'Extracurricular_Activities', 'Learning_Disabilities', 'School_Type', 'Internet_Access']
```

# Standardization

Since some features are more likely to be numerically large, standardization was utilized to make sure the model does not assign greater weights to those features even though they are not truly significant.

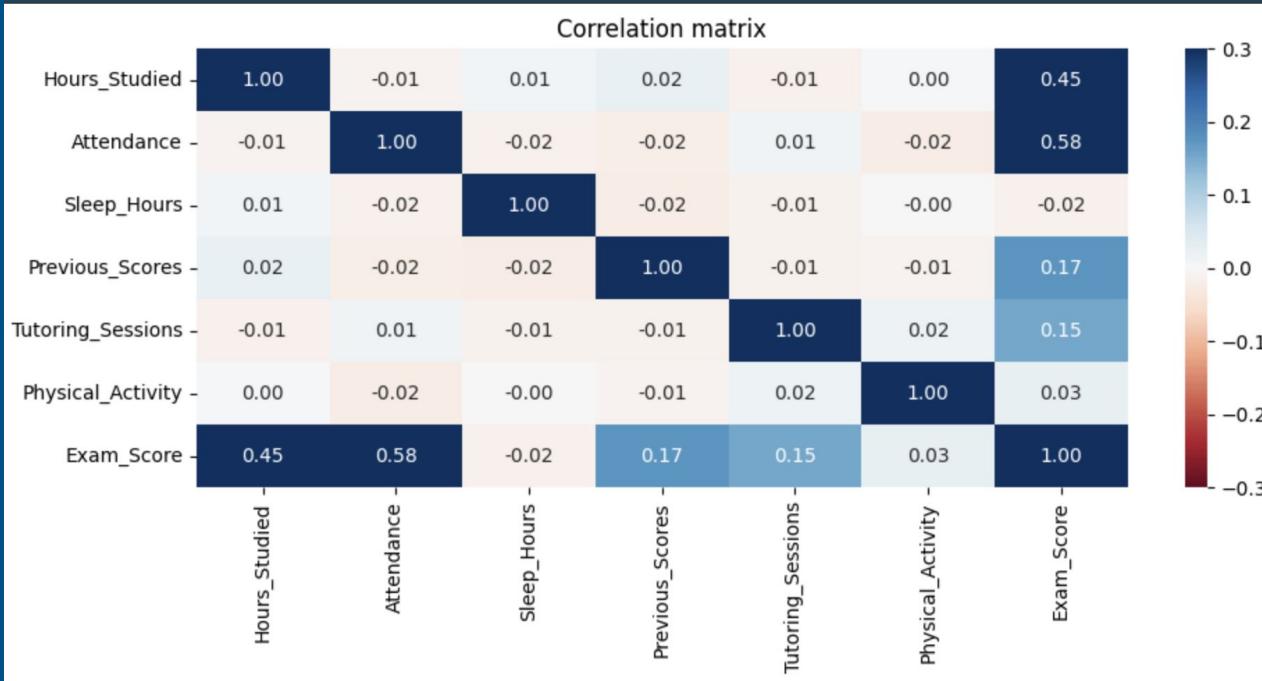
For instance, exam scores are in the range of 0-100% but the average student only sleeps 0-10 hours. But, since exam scores are greater than hours of sleep, the model may inherently assign higher weights to exam scores. Therefore, normalization is required prior to creating the machine learning models.

```
X = data[['Attendance', 'Hours_Studied']]
y = data['Exam_Score']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=34)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

# Exploratory Data Analysis



Based off of the results of the correlation matrix we can reason that none of the numeric independent variable are strongly correlated with the dependent variable, Exam Score. At best the they are weakly correlated with Exam Score, notably Hours Studied, Attendance, Previous Scores, and Tutoring Sessions. There also seems to be even less correlation ( $-0.02 < \text{corr} < 0.03$ ) between the independent variables.

# Exploratory Data Analysis

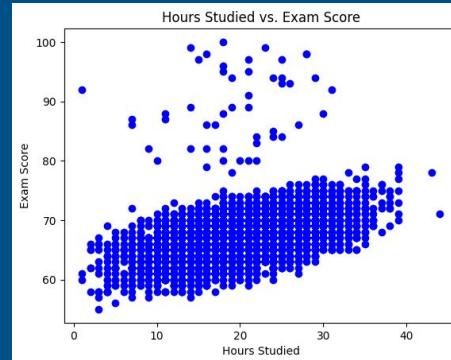
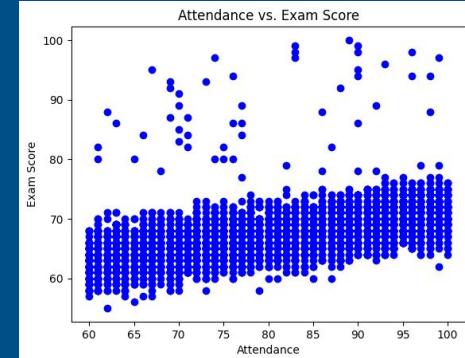
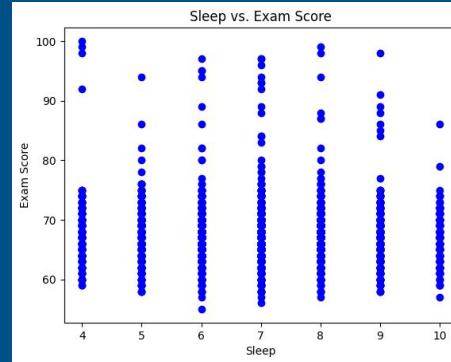
A Spearman correlation matrix was additionally produced alongside the previous Pearson's correlation matrix. Notably, correlation between Attendance and Exam Score increased from approximately 0.58 to 0.67 indicating an even stronger positive relationship. Overall both correlation matrices are fairly equal indicating a monotonic, linear relationship between the variables. With this we can hypothesize that linear regression would be the best suited model to evaluate our problem.

	Hours_Studied	Attendance	Sleep_Hours	Previous_Scores	Tutoring_Sessions	Physical_Activity	Exam_Score	Parental_Involvement	Access_to_Resources	Tutoring_Sessions	Family_Income	Teacher_Quality	Parental_Education_Level
Hours_Studied	1.000000	-0.010127	0.011360	0.023859	-0.013276	-0.002656	0.480775	-0.018708	-0.008706	-0.013276	-0.002086	-0.004056	-0.008144
Attendance	-0.010127	1.000000	-0.012433	-0.020362	0.014365	-0.024196	0.672235	-0.008369	-0.011321	0.014365	-0.014421	-0.000326	0.024972
Sleep_Hours	0.011360	-0.012433	1.000000	-0.022138	-0.005913	0.000875	-0.007403	-0.006694	-0.012984	-0.005913	-0.018209	0.003487	0.010545
Previous_Scores	0.023859	-0.020362	-0.022138	1.000000	-0.018523	-0.008234	0.191679	-0.020525	0.023294	-0.018523	-0.013048	-0.001215	-0.012948
Tutoring_Sessions	-0.013276	0.014365	-0.005913	-0.018523	1.000000	0.007500	0.163552	0.000687	-0.013908	1.000000	0.003253	0.001282	0.005741
Physical_Activity	-0.002656	-0.024196	0.000875	-0.008234	0.007500	1.000000	0.029148	-0.004107	-0.010583	0.007500	-0.019170	-0.013583	-0.029173
Exam_Score	0.480775	0.672235	-0.007403	0.191679	0.163552	0.029148	1.000000	0.172663	0.187363	0.163552	0.097753	0.082021	0.115172
Parental_Involvement	-0.018708	-0.008369	-0.006694	-0.020525	0.000687	-0.004107	0.172663	1.000000	-0.024933	0.000687	0.012875	0.011744	-0.006115
Access_to_Resources	-0.008706	-0.011321	-0.012984	0.023294	-0.013908	-0.010583	0.187363	-0.024933	1.000000	-0.013908	-0.006929	-0.010447	-0.005434
Tutoring_Sessions	-0.013276	0.014365	-0.005913	-0.018523	1.000000	0.007500	0.163552	0.000687	-0.013908	1.000000	0.003253	0.001282	0.005741
Family_Income	-0.002086	-0.014421	-0.018209	-0.013048	0.003253	-0.019170	0.097753	0.012875	-0.006929	0.003253	1.000000	-0.005644	0.000171
Teacher_Quality	-0.004056	-0.000326	0.003487	-0.001215	0.001282	-0.013583	0.082021	0.011744	-0.010447	0.001282	-0.005644	1.000000	0.000032
Parental_Education_Level	-0.008144	0.024972	0.010545	-0.012948	0.005741	-0.029173	0.115172	-0.006115	-0.005434	0.005741	0.000171	0.000032	1.000000

# Exploratory Data Analysis

Analyzing these scatterplots we can infer about the relationships between the dependent variable, Exam Score and the independent variables: Sleep Hours, Attendance and Hours Studied.

1. Sleep v.s. Exam Score
  - a. Sleep is weakly negatively correlated with Exam Score
2. Attendance v.s. Exam Score
  - a. Attendance is weakly positively correlated with Exam Score
3. Hours Studied v.s. Exam Score
  - a. Hours Studied is weakly positively correlated with Exam Score.



# Modeling

Based off the EDA results, linear regression was the model that would likely lead to optimal performance. However, lasso regularization was explored as it may perform better than linear regression.

With the result of both models, we concurred that simple linear regression was the better modeling method due to the higher R<sup>2</sup> score meaning that the model accounted for more variation compared to lasso regularization. Simple linear regression also had lower error scores indicating it was a better suited model.

## Lasso Regularization

```
Exam_Score = 67.24337623012869
+ (1.7608223293174632 * Hours_Studied)
+ (2.2768102042056517 * Attendance)
+ (0.696974458448327 * Parental_Involvement)
+ (0.7303808626246465 * Access_to_Resources)
+ (0.2859954470186946 * Extracurricular_Activities)
+ (-0.0018738143188387014 * Sleep_Hours)
+ (0.6891865544554763 * Previous_Scores)
+ (0.37941202112766936 * Motivation_Level)
+ (0.2470870526089079 * Internet_Access)
+ (0.6048415914373912 * Tutoring_Sessions)
+ (0.3616106305751394 * Family_Income)
+ (0.29839240243897036 * Teacher_Quality)
+ (0.38167667305240194 * Peer_Influence)
+ (0.17671446515883826 * Physical_Activity)
+ (-0.23183144394379 * Learning_Disabilities)
+ (0.3920799352739844 * Parental_Education_Level)
+ (-0.31044639693718895 * Distance_from_Home)
+ (-0.025622592920181996 * Gender)
+ (-0.002497806291395969 * School_Type_Private)
+ (0.002497806291431445 * School_Type_Public)
```

Training Mean Squared Error: 3.982190730177149  
 Training Root Mean Squared Error: 1.995542715698451  
 Training R<sup>2</sup> Score: 0.7311799305740225

Testing Mean Squared Error: 4.191798699868268  
 Testing Root Mean Squared Error: 2.047388263097224  
 Test R<sup>2</sup> Score: 0.7305543179365184

## Simple Linear Regression

```
Exam_Score = 67.24337623012869
+ (1.760856035987102 * Hours_Studied)
+ (2.276854359416088 * Attendance)
+ (0.6969896043147007 * Parental_Involvement)
+ (0.7303954474804775 * Access_to_Resources)
+ (0.2860005172319154 * Extracurricular_Activities)
+ (-0.0018726694375620356 * Sleep_Hours)
+ (0.6892008320176293 * Previous_Scores)
+ (0.37941991364159744 * Motivation_Level)
+ (0.24709244666291186 * Internet_Access)
+ (0.6048528555123398 * Tutoring_Sessions)
+ (0.36161862520707916 * Family_Income)
+ (0.29839816787003526 * Teacher_Quality)
+ (0.3816839739874137 * Peer_Influence)
+ (0.17672007926199798 * Physical_Activity)
+ (-0.23183490181314587 * Learning_Disabilities)
+ (0.39208700192225704 * Parental_Education_Level)
+ (-0.3104522890079516 * Distance_from_Home)
+ (-0.025624056320901624 * Gender)
+ (-0.0024983748200244627 * School_Type_Private)
+ (0.0024983748200249067 * School_Type_Public)
```

Training Mean Squared Error: 3.982190726063726  
 Training Root Mean Squared Error: 1.9955427146677984  
 Training R<sup>2</sup> Score: 0.7311799308517015

Testing Mean Squared Error: 4.191783933205914  
 Testing Root Mean Squared Error: 2.0473846568746956  
 Test R<sup>2</sup> Score: 0.7305552671265187

# Model Evaluation

## Simple Linear Regression

### Hours Studied & Attendance

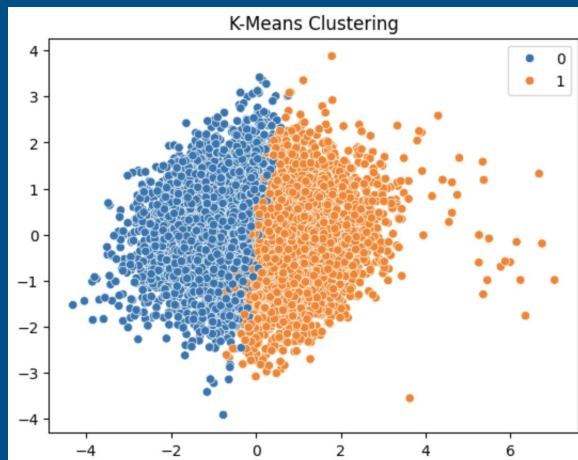
```
Exam_Score = 67.24337623012869  
+ (2.248330188157404 * Attendance)  
+ (1.7557330920923804 * Hours_Studied)  
  
Training Mean Squared Error: 6.732863491939162  
Training Root Mean Squared Error: 2.594776193034606  
Training R^2 Score: 0.5454941880048525  
  
Testing Mean Squared Error: 7.190366506997066  
Testing Root Mean Squared Error: 2.681485876710348  
Test R^2 Score: 0.5378086242965047
```

### Hours Studied, Attendance, Access to Resources & Parental Involvement

```
Exam_Score = 67.24337623012869  
+ (1.7599873519430944 * Hours_Studied)  
+ (2.263220980608788 * Attendance)  
+ (0.728349010628579 * Access_to_Resources)  
+ (0.6736646823016736 * Parental_Involvement)  
  
Training Mean Squared Error: 5.782669223780852  
Training Root Mean Squared Error: 2.404718117322871  
Training R^2 Score: 0.6096375971084942  
  
Testing Mean Squared Error: 6.209498695407191  
Testing Root Mean Squared Error: 2.491886573543666  
Test R^2 Score: 0.6008580728581099
```

# Model Evaluation

## K-Means Clustering



Additionally, k-means clustering was implemented to show the different patterns and characteristics of clusters and their impact on Exam Score.

Based on the results cluster 0 performed better than cluster 1 on the final exam. The reasons for this is because of their higher attendance and study hours. Hours slept had little to no effect on clustering.

	Exam_Score	Hours_Studied	Attendance	Sleep_Hours
cluster				
0	70.018700	21.708716	89.183201	7.008875
1	64.681542	18.388583	71.556071	7.047812

# Conclusion

Overall, it is clear that the best predictors for a student's academic success is attendance and hours studied. This can further be complimented by the student's access to resources and parental involvement. Teachers and other academic staff can use these metrics to better understand if a student needs extra support. Early intervention can greatly alter the course of a student's academic performance.