COMP S460F

Advanced Topics in Data Mining

Group 12

Factors Influencing Student Performance: A Comprehensive Data Analysis Approach

Group member	Student ID
Yip Tin Hang	13258976
Choy Hei Kiu Max	13027559
Lu Yuk Tong	13439007

Abstract

This study investigates the factors influencing student performance using a dataset sourced from Kaggle, comprising 20 columns and 6,607 records. The research aims to provide insights into various aspects such as attendance, study habits, and parental involvement that significantly impact academic achievement. The methodology includes data preprocessing steps, followed by employing stepwise regression techniques to predict exam scores. Additionally, polynomial fitting is utilized for enhanced predictions, and the results are visualized through various plots. This approach not only facilitates a detailed understanding of the underlying factors affecting student performance but also serves as a foundation for developing predictive models which can be expanded to broader educational applications.

Content

1. Introduction

In this project, we aim to analyze the factors influencing student performance using a dataset sourced from Kaggle, which contains 20 columns and 6,607 records. This dataset provides valuable insights into various aspects such as attendance, study habits, and parental involvement that significantly impact academic achievement. Our methodology includes data preprocessing steps, followed by employing stepwise regression techniques to predict exam scores. Additionally, we will apply polynomial fitting for enhanced predictions and visualize the results through various plots.

Understanding these factors is crucial for educators and policymakers to develop strategies that can improve student outcomes. By identifying key influences on academic performance, this study aims to contribute to the broader field of educational research and provide a foundation for future predictive models that can be applied in diverse educational settings.

2. Problem description

<table>
<tr><td>Ordinary Least Squares regression model</td><td>Polynomial Regression Model</td></tr>
<tr><td>$model : y = X\beta + \epsilon$</td><td>$y = \beta_0 + \beta_1 X + \beta_2 X^2$</td></tr>
<tr><td colspan="2">Ridge Regression</td></tr>
<tr><td colspan="2">$Min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})) + \lambda \sum_{p=1}^{2} \beta_p^2$</td></tr>
<tr><td colspan="2">Lasso Regression</td></tr>
<tr><td colspan="2">$Min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}))^2 + \lambda \sum_{p=1}^{2} \lvert \beta_p \rvert$</td></tr>
<tr><td>Mean Squared Error</td><td>Mean Absolute Error</td></tr>
</table>

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
Root Mean Squared Error	R-squared
$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$
Adjusted R-squared	
$Adjusted R^2 = 1 - (1 - R^2) * \frac{n-1}{n-p-1}$	

3. Methodology

3.1 Dataset / Data Collection & Pre-processing

Online Dataset: our project is based on the online dataset provide by Kaggle about student performance factor which contains 20 columns and 6607 records, it provides a comprehensive overview of several factors affecting student performance in exams. It contains data on attendance, study habits, parental participation, and other factors that affect academic achievement.

Before we application these data, we will do the following pre-processing: data clean and encode the non-numeric data.

Step	**Procedure**
1. Load dataset	Load the dataset of csv file
2. Remove the Duplicates	Remove duplicates data
3. Encode data	Encode the categorical variables into numeric columns.

3.2 Stepwise Regression (Linear Regression)

In this part, we use stepwise regression techniques to predict exam scores based on various features. Starting with backward elimination, we remove insignificant features. Then, we apply Ridge Regression to reduce overfitting and Lasso Regression for feature selection. Finally, we evaluate model performance using metrics like Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared to assess accuracy and reliability.

Step	**Procedure**
1. Feature Selection	Separate features from the target variable. X is the other column without 'Exam_Scores'. y is assigned the 'Exam_Score' column.
2. Backward Elimination	Define a function to iteratively remove non-significant features based on p-values until all remaining features are significant (p-value ≤ 0.05).
3. Fit Model	Inside the backward elimination function, fit an Ordinary Least Squares model with the features and target variable.
4. Check P-value	Extract p-values of the features and identify the maximum p-value. If it exceeds 0.05, remove the corresponding feature and refit the model.
5. Final Model	Return the final model after all non-significant features are removed.
6. Ridge Regression	Create and fit a Ridge Regression model using the training data. Make predictions on the test data.
7. Evaluate Ridge Model	Calculate the R^2 score for Ridge predictions.
8. Lasso Regression	Create and fit a Lasso Regression model using the same training data.
9. Evaluate Lasso Model	Calculate the R^2 score for Lasso predictions.

10. Calculate Errors	Compute MSE for both Ridge and Lasso models using y_test and predicted values.
11. Comprehensive Evaluation	Calculate additional evaluation metrics (MAE, RMSE, R^2) for one of the models (e.g., Ridge).

3.3 Polynomial Fitting (Curve Fitting)

In this part, we process student performance data, remove duplicates, and encode categorical variables. It splits the data into training and testing sets, standardizes the features, and fits a polynomial Ridge regression model. The model predicts exam scores, and performance metrics (MSE, RMSE, MAE, R^2) are calculated. Finally, it visualizes the results in a 3D plot of hours studied, attendance, and exam scores.

Step	**Procedure**
1. Define Features	Set features (X) and target (y).
2. Split Data	Split into training (85%) and testing (15%).
3. Standardize Features	Scale the features.
4. Fit Model	Create polynomial features and fit a Ridge regression model.
5. Predict Scores	Generate predictions on the test set.
6. Calculate Metrics	Compute MSE, RMSE, MAE, and R^2.
7. Visualize	Plot the results in a 3D graph.

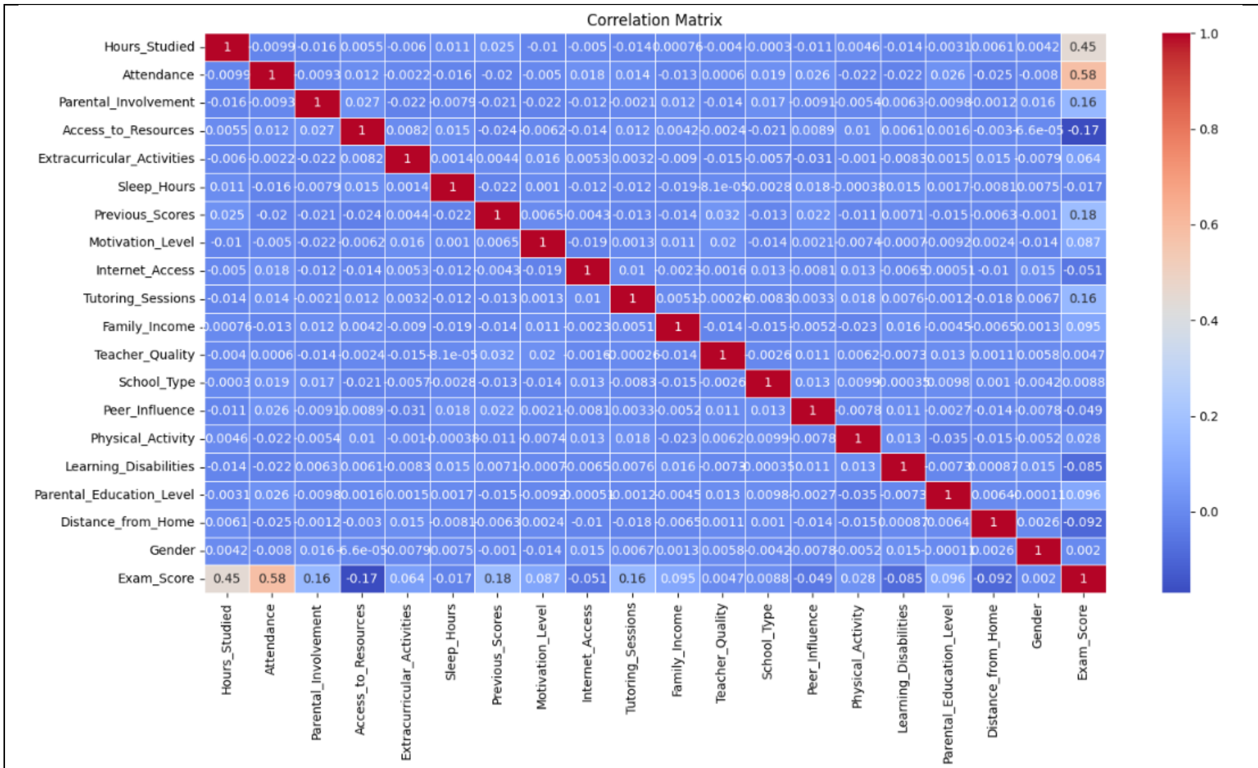
3.4 Visualization

Step	**Procedure**
1. Heatmap	Using Correlation Heatmap to plot a heatmap to visualize correlations among numerical features
2. Regression Line	Using Regression Line to plot the relationship between predictors and the target variable.
3. IQR	Visualize the relationship between Exam Scores and Attendance

4. Volin Plot	Use a violin plot to displays the distribution and variability of exam score within each category of Gender, revealing trends and group differences.
5. Bubble Plot	Create a bubble plot to visualize the relationship between multiple features. To demonstrates pairwise comparisons with visual cues for additional attributes.

4. Analysis

4.1 Heatmap - Correlation Matrix



The correlation matrix visually represents the relationships between various predictors and the target variable, Exam_Score, with values ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

The Strong Positive Correlations with Exam_Score: Hours_Studied (0.45), Attendance (0.58), Parental_Involvement (0.16).

The Negative Correlations with Exam_Score: Access_to_Resources (-0.17), Learning_Disabilities (-0.09).

4.2 Stepwise regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Exam_Score   R-squared:                       0.702
Model:                            OLS   Adj. R-squared:                  0.701
Method:                 Least Squares   F-statistic:                     827.9
Date:                Wed, 20 Nov 2024   Prob (F-statistic):               0.00
Time:                        15:26:37   Log-Likelihood:                -11521.
No. Observations:                5285   AIC:                         2.307e+04
Df Residuals:                    5269   BIC:                         2.318e+04
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         39.3910      0.312    126.281      0.000      38.780      40.003
Hours_Studied                  0.2919      0.005     59.265      0.000       0.282       0.302
Attendance                     0.1992      0.003     77.632      0.000       0.194       0.204
Parental_Involvement           1.0220      0.042     24.069      0.000       0.939       1.105
Access_to_Resources           -1.0302      0.042    -24.342      0.000      -1.113      -0.947
Extracurricular_Activities     0.5855      0.060      9.725      0.000       0.467       0.704
Previous_Scores                0.0492      0.002     23.991      0.000       0.045       0.053
Motivation_Level               0.5161      0.043     12.137      0.000       0.433       0.599
Internet_Access               -0.9440      0.110     -8.574      0.000      -1.160      -0.728
Tutoring_Sessions              0.5073      0.024     21.203      0.000       0.460       0.554
Family_Income                  0.5466      0.040     13.751      0.000       0.469       0.625
Peer_Influence                -0.2487      0.033     -7.489      0.000      -0.314      -0.184
Physical_Activity              0.1815      0.029      6.350      0.000       0.125       0.238
Learning_Disabilities         -0.8858      0.094     -9.385      0.000      -1.071      -0.701
Parental_Education_Level       0.4242      0.036     11.749      0.000       0.353       0.495
Distance_from_Home            -0.4249      0.041    -10.270      0.000      -0.506      -0.344
==============================================================================
Omnibus:                     8369.932   Durbin-Watson:                   1.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3242972.531
Skew:                          10.433   Prob(JB):                         0.00
Kurtosis:                     122.547   Cond. No.                     1.19e+03
==============================================================================
```

The OLS regression analysis shows that the model explains 70.2% of the variance in Exam Scores. Key findings include:

- **Positive Influences**: Hours Studied, Attendance, Parental Involvement, and Motivation Level significantly boost Exam Scores.
- **Negative Influences**: Access to Resources and Learning Disabilities negatively affect scores.
- **Drop out feature:** School_Type, Gender, Teacher_quality, Sleep_Hours.

All predictors are statistically significant ($p < 0.001$), indicating their importance in student performance.

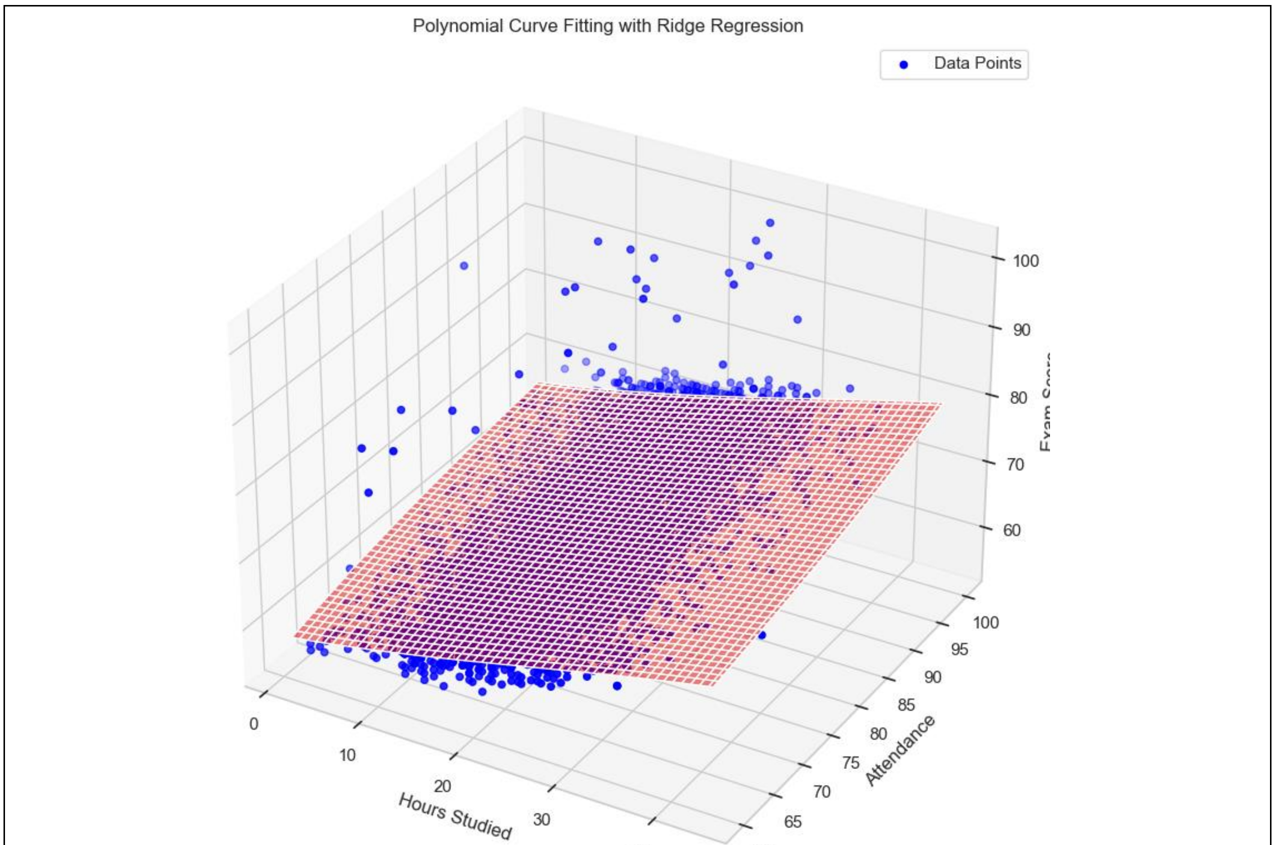
Model Evaluation Metrics: R-squared: 0.7570 Mean Absolute Error (MAE): 0.5911 Mean Squared Error (MSE): 3.4348 Root Mean Squared Error (RMSE): 1.8533	• **MAE(0.59)**: Indicates that, on average, the model's predictions are off by 0.59 units, which is relatively low. • **MSE(3.43)** and **RMSE(1.85)**: Both metrics suggest that the model has a moderate level of error, with RMSE providing a more interpretable scale since it is in the same units as the target variable. • **R^2 (0.757)**: This value indicates that approximately 75.7% of the variance in the target variable is explained by the model, suggesting a good fit.
Lasso Model Evaluation Metrics: R-squared: 0.6179 Ridge Model Evaluation Metrics: R-squared: 0.7570	**Ridge R^2(0.76)** vs. **Lasso R^2(0.74):** Ridge has a higher R^2 score, confirming its superior performance in explaining variance.
Lasso MSE: 5.4006 Ridge MSE: 3.4347	**Ridge MSE(3.43)** vs. **Lasso MSE(5.40):** Ridge performs slightly better in terms of MSE, indicating better predictive accuracy.

4.3 Curve Fitting (Polynomial fitting - Cross-validation)

Features such as 'Hours_Studied' and 'Attendance' are selected as predictors for the target variable 'Exam_Score'. The data is split into training and testing sets, with features standardized for better model performance. A polynomial transformation (degree 3) is applied to the features, followed by Ridge regression with cross-validation to determine the optimal alpha value.

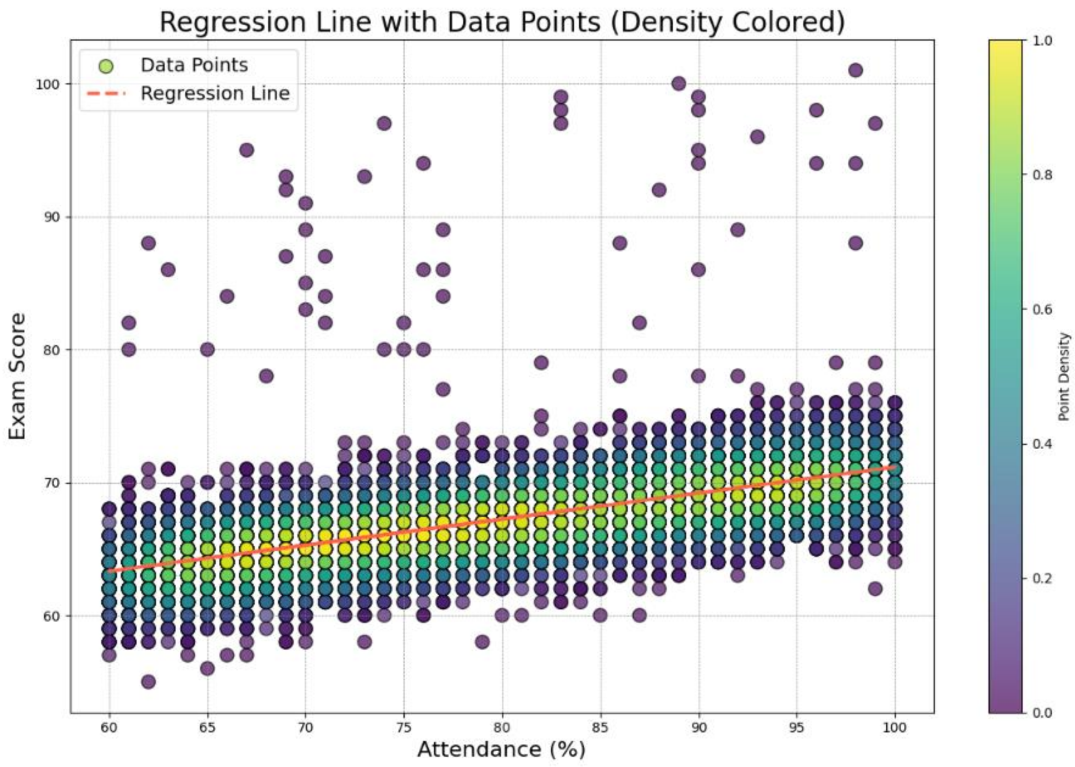
```
Mean Squared Error (MSE): 5.61
Root Mean Squared Error (RMSE): 2.37
Mean Absolute Error (MAE): 1.48
R-squared (R²): 0.61
Adjusted R-squared: 0.61

Cross-Validation Results:
Alpha: 0.000001, Cross-Validated Score: 4.1474
Alpha: 0.000010, Cross-Validated Score: 0.6437
Alpha: 0.000100, Cross-Validated Score: 3.9686
Alpha: 0.001000, Cross-Validated Score: 6.5778
Alpha: 0.010000, Cross-Validated Score: 0.5043
Alpha: 0.100000, Cross-Validated Score: 16.1124
Alpha: 1.000000, Cross-Validated Score: 0.0487
Alpha: 10.000000, Cross-Validated Score: 1.8151
Alpha: 100.000000, Cross-Validated Score: 2.8571
Alpha: 1000.000000, Cross-Validated Score: 0.7666
Alpha: 10000.000000, Cross-Validated Score: 15.3781
Alpha: 100000.000000, Cross-Validated Score: 4.3976
Alpha: 1000000.000000, Cross-Validated Score: 2.4470
```

- **RMSE(2.37):** This means the model's predictions are, on average, off by about 2.37 points, indicating a moderate level of error.
- **MAE(1.48):** The model's predictions are typically within 1.48 points of the actual scores, which is easier to interpret.
- **R^2 Value(0.61):** This indicates that 61% of the variability in exam scores can be explained by hours studied and attendance, suggesting a good fit.
- **Cross-Validation Results:** The optimal alpha is around 0.1, with a cross-validated score of 16.1124, indicating minimal error.



Blue points represent the original data points in the 'Hours Studied' vs. 'Attendance' vs. 'Exam Score' space. The red surface shows the polynomial fit generated by the Ridge regression model, indicating how the predicted 'Exam Score' varies with changes in 'Hours Studied' and 'Attendance'. This visual aid helps in understanding the complexity and behavior of the relationship between the input features and the target variable, as well as assessing the quality of the polynomial approximation.

4.4 Regression Line



Axes:

- X-axis: Attendance percentage
- Y-axis: Exam scores

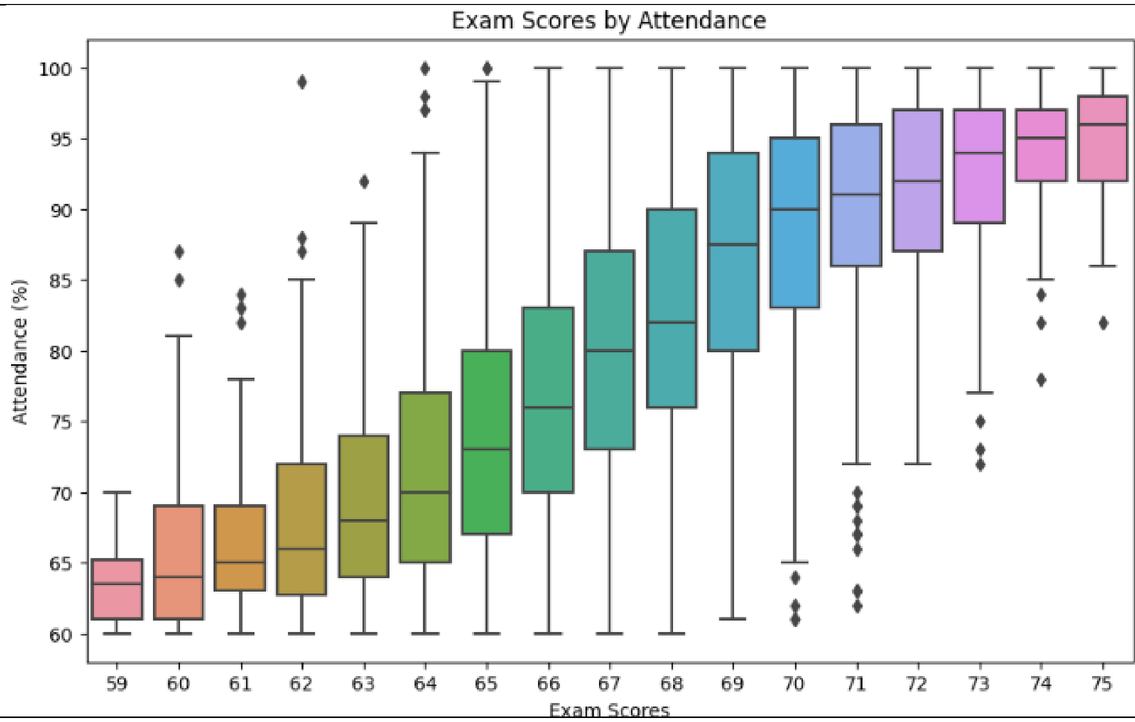
Color Coding: Points are colored based on density, with brighter shades indicating areas with higher concentrations of data points.

Regression Line: A dashed linear regression line (in tomato color) indicates a positive correlation: As attendance increases, exam scores tend to increase.

Variability: Data points do not perfectly align with the regression line, suggesting other factors may influence exam scores.

The analysis suggests a general trend where higher attendance is associated with better exam performance. And the regression model basically conforms to the data point trend.

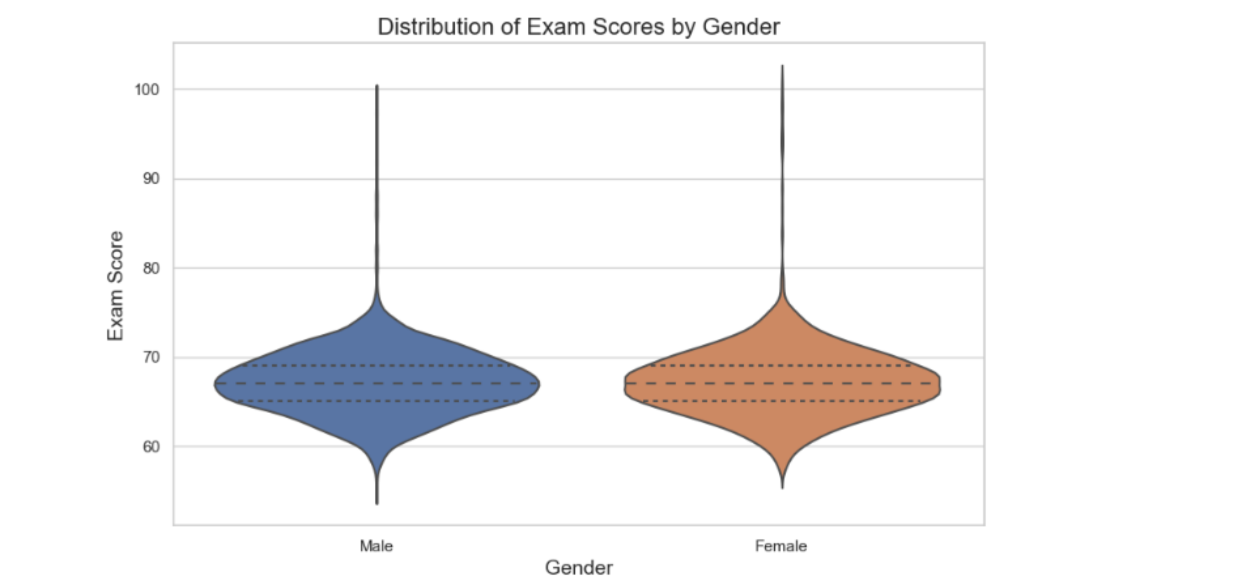
4.5 IQR



This plot analyzes exam scores and attendance rates across subjects, excluding outliers for clarity. Each box represents the distribution of scores, with the height indicating the interquartile range from the 25th to 75th percentile. The central line marks the median score, while whiskers extend to show the range of non-outlier scores.

The plot reveals variations in exam performance linked to attendance levels, suggesting that higher attendance tends to correlate with better exam scores.

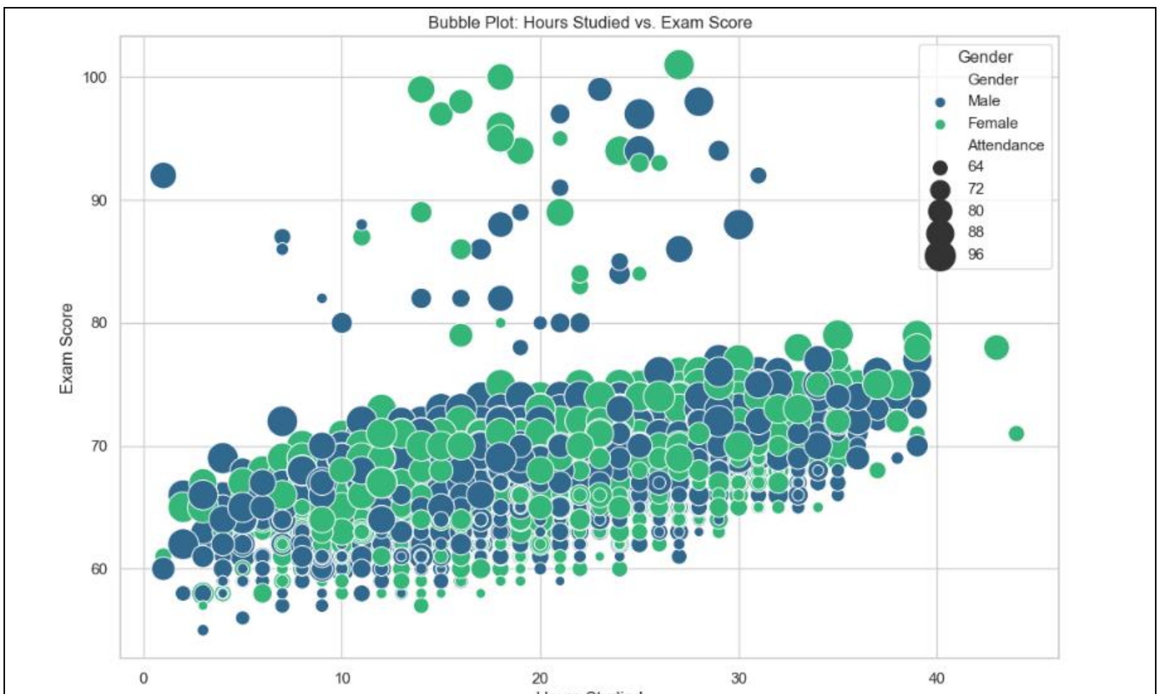
4.6 Violin Plots



This plot illustrates the distribution of exam scores by gender, showing the scores (y-axis) for male and female students (x-axis).

Both genders have similar median scores; however, females display greater variability and tend to outperform males at the higher end of the score distribution.

4.7 Bubble Plot



The bubble plot illustrates the relationship between hours studied (x-axis), attendance (bubble size), and exam scores.

It demonstrates a strong positive correlation between hours studied and attendance with exam performance. The prevalence of small bubbles in the lower section indicates that attendance significantly impacts exam scores. Additionally, it highlights gender differences, showing that females slightly outperform males at higher score levels.

5. Conclusion

This project utilized a comprehensive dataset from Kaggle to analyze factors influencing student performance, focusing on elements such as attendance, study habits, and parental involvement. Through rigorous data preprocessing, stepwise regression techniques, and polynomial fitting, we aimed to predict exam scores accurately.

Our results indicated significant positive correlations between exam scores and factors like hours studied, attendance, and parental involvement. Conversely, access to resources and learning disabilities showed negative correlations. The OLS regression model explained 75.7% of the variance in exam scores, with all predictors being statistically significant.

Visualizations employed, including heatmaps and violin plots, provided clear insights into the relationships between variables, further validating our findings. This analysis not only offers deeper understanding of student performance determinants but also paves the way for developing robust predictive models with potential applications in educational policymaking.

However, this study has limitations. The dataset is sourced from Kaggle and may not represent a global or diverse population. Additionally, the model's performance could be influenced by unobserved confounding factors.

Future work can focus on expanding the dataset to include more diverse populations, incorporating additional variables like socio-economic status, and exploring other machine learning techniques for better predictive accuracy.

6. Appendix

Project Code: https://github.com/LutherYTT/COMPS460F-Factors-Influencing-Student-Performance

7. Reference

Kaggle. (2023, January 10). Student performance factors dataset. Kaggle Datasets. https://www.kaggle.com/datasets/lainguyn123/student-performance-factors