Detailed Analysis and Report

## **Student Performance Analysis Report**

### **1. Introduction**

This report analyses student performance using traditional (mean) and quantile multiple regression techniques. The dataset used for this analysis is "Student Performance - Multiple Linear Regression" from Kaggle.

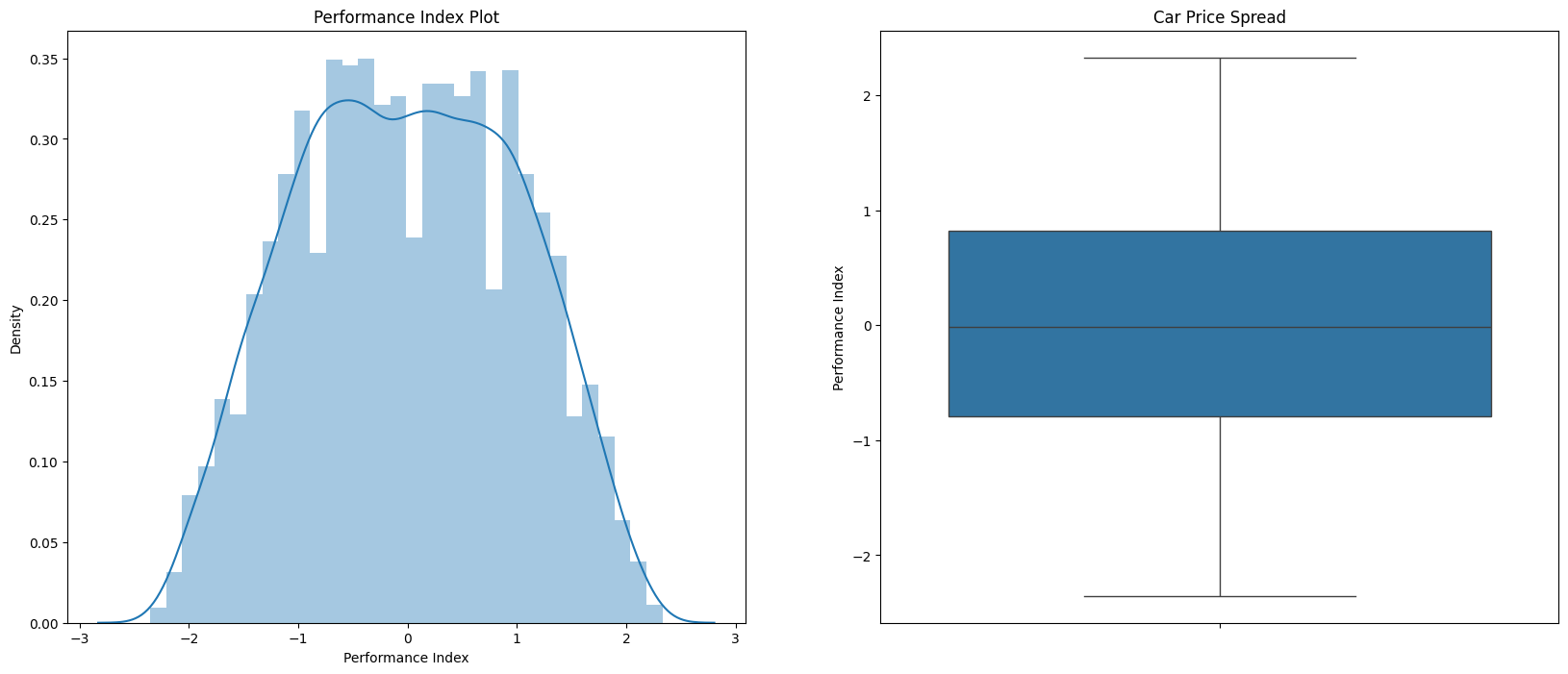
### **2. Exploratory Data Analysis (EDA)**

#### **2.1 Descriptive Statistics**

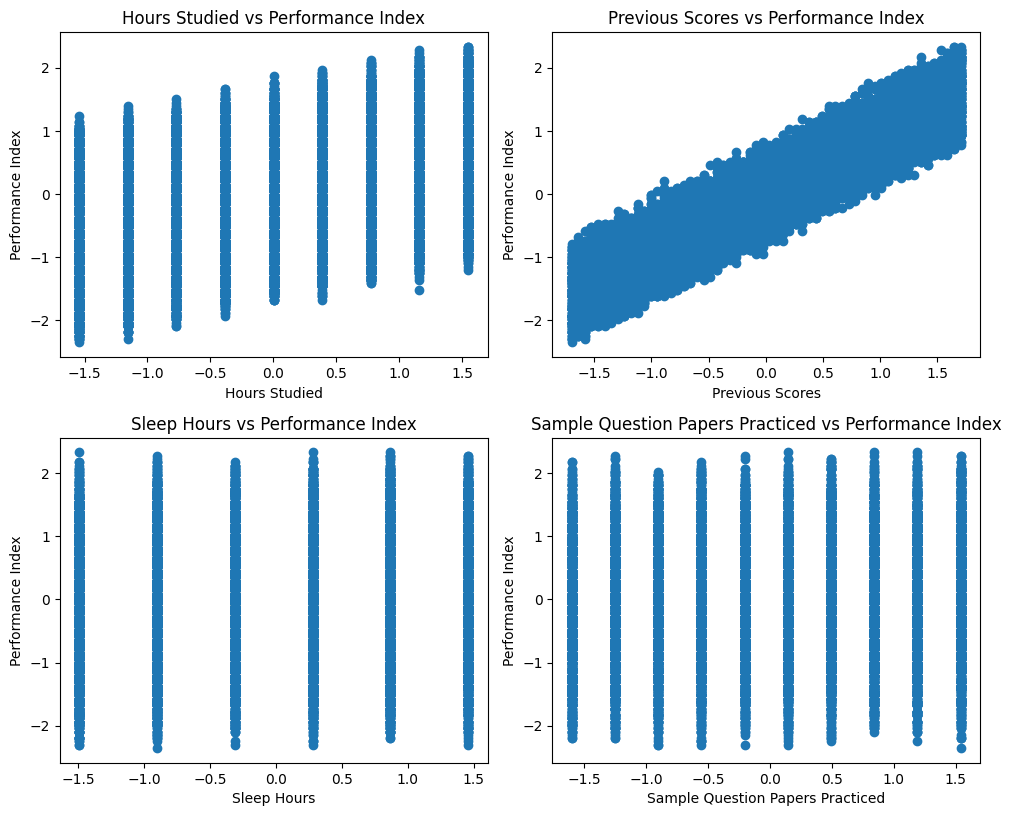
Descriptive statistics were computed for all features in the dataset to understand the central tendency, dispersion, and shape of the distribution.

#### **2.2 Data Visualization**

* **Distribution and Box Plot:**
  + Plotted the distribution and box plot for 'Performance Index'.



* **Scatter Plots:**
  + Generated scatter plots for 'Hours Studied', 'Previous Scores', 'Sleep Hours', and 'Sample Question Papers Practiced' against 'Performance Index'.



### **3. Data Preprocessing**

#### **3.1 Handling Missing Values**

* No missing values were detected in the dataset.

#### **3.2 Feature Scaling**

* Numerical features were standardized using StandardScaler.

#### **3.3 Categorical Encoding**

* Categorical features were encoded using OneHotEncoder.

#### **3.4 Concatenation**

* Combined the scaled numerical features and encoded categorical features into a single Data Frame.

### 

### 

### 

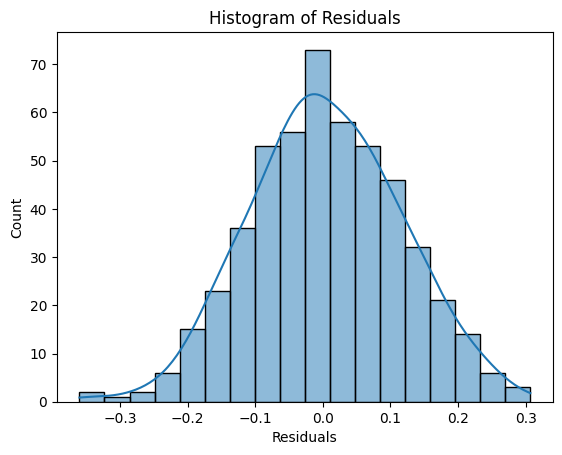
### **4. Traditional Multiple Linear Regression**

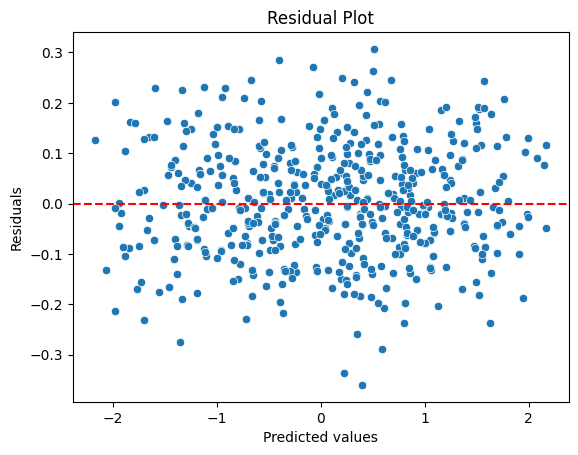
#### **4.1 Model Training**

* Trained a linear regression model using the pre-processed training data.

#### **4.2 Model Evaluation**

* **Metrics:**
  + R² Score:
  + Training R-squared: 0.9886771025449246
  + Testing R-squared: 0.9876600290218607
  + Mean Squared Error (MSE):
  + Training Mean Squared Error: 0.011313279896213663
  + Testing Mean Squared Error: 0.012339785454698686
* **Residual Analysis:**
  + Plotted residuals and histograms of residuals to evaluate the model fit.





### 

### **5. Quantile Regression**

#### **5.1 Model Training**

* Trained quantile regression models for the 0.25, 0.5, and 0.75 quantiles.

#### **5.2 Model Evaluation**

* **Metrics:**
  + Mean Squared Error (MSE) for each quantile:

Quantile 0.25 - Training Mean Squared Error: 0.01642020050280117

Quantile 0.25 - Testing Mean Squared Error: 0.01806830780075373

Quantile 0.5 - Training Mean Squared Error: 0.011318338729604577

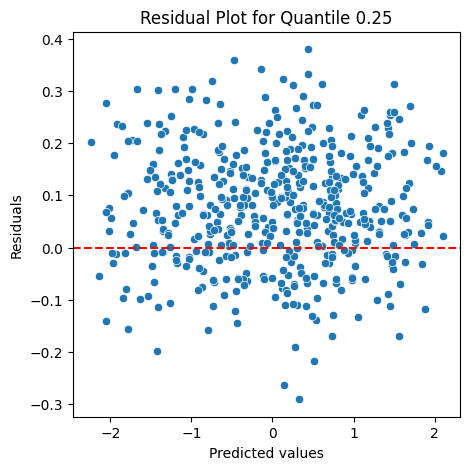
Quantile 0.5 - Testing Mean Squared Error: 0.012352043647055996

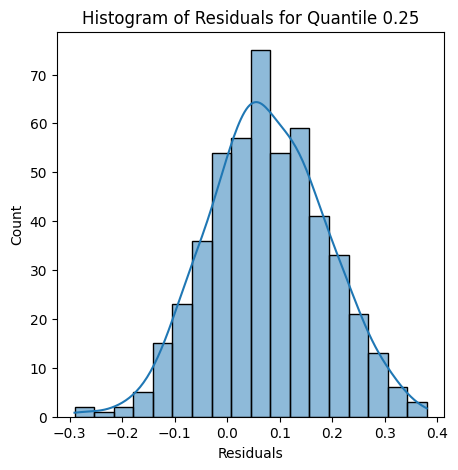
Quantile 0.75 - Training Mean Squared Error: 0.0162151990473928

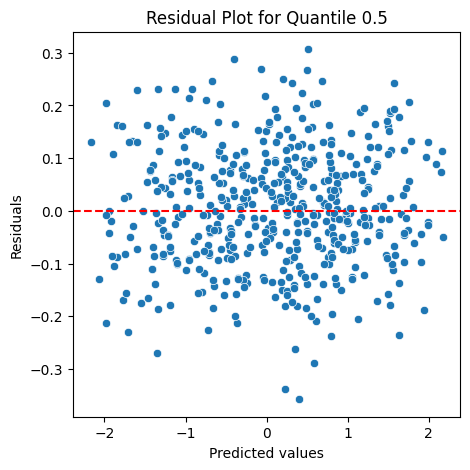
Quantile 0.75 - Testing Mean Squared Error: 0.016628480423824775

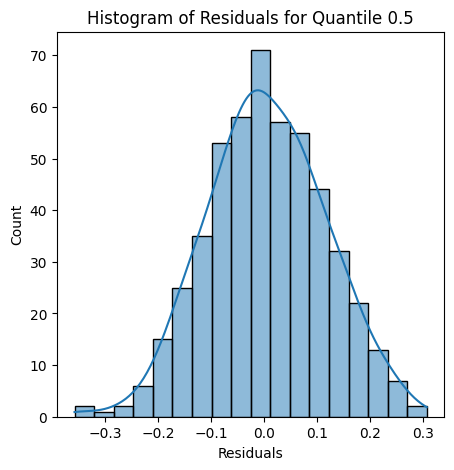
* **Residual Analysis:**
  + Plotted residuals and histograms of residuals for each quantile.

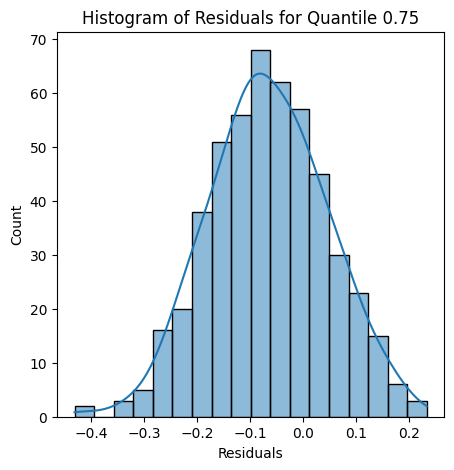
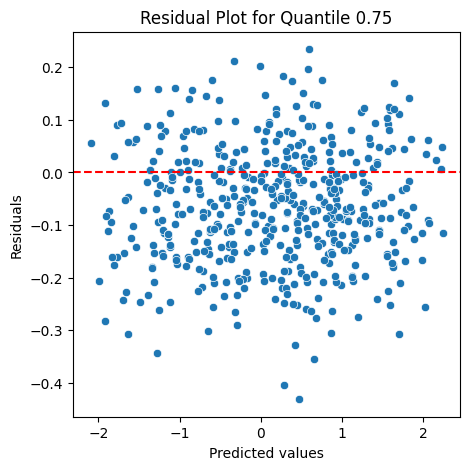
**0.25 Quantile:**

****

****

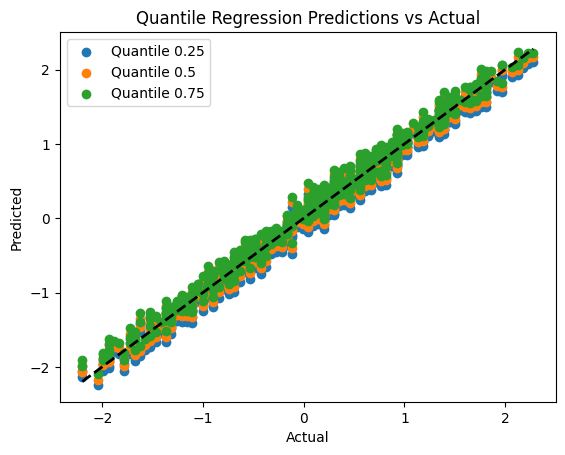
****

****

****

#### **5.3 Visualization**

* Plotted predictions vs actual values for each quantile.



### **6. Comparison and Conclusion**

#### **6.1 Model Performance**

**Traditional Multiple Linear Regression:**

**R² Score:**

**Training R-squared:** 0.9887

Indicates that approximately 98.87% of the variance in the performance index is explained by the model during training.

**Testing R-squared:** 0.9877

Indicates that approximately 98.77% of the variance in the performance index is explained by the model during testing.

**Mean Squared Error (MSE):**

**Training MSE:** 0.0113

A low MSE value indicates a small average squared difference between the actual and predicted values during training.

**Testing MSE:** 0.0123

Similarly low MSE value on the test set indicates that the model generalizes well to unseen data.

**Quantile Regression:**

**0.25 Quantile:**

**Training MSE:** 0.0164

Higher than the traditional model's MSE, suggesting less accuracy at the lower quantile.

**Testing MSE:** 0.0181

Indicates the model is less accurate when predicting lower quantiles compared to the traditional model.

**0.5 Quantile (Median):**

**Training MSE:** 0.0113

Very close to the traditional model's MSE, indicating similar accuracy for the median.

**Testing MSE:** 0.0124

Also close to the traditional model, reinforcing the model's effectiveness in predicting the median performance index.

**0.75 Quantile:**

**Training MSE:** 0.0162

Higher than the traditional model's MSE, suggesting less accuracy at the upper quantile.

**Testing MSE:** 0.0166

Indicates the model is less accurate when predicting upper quantiles compared to the traditional model.

#### 

#### 

#### **6.2 Insights**

**Model Fit:**

The traditional multiple linear regression model demonstrates a strong fit with high R² scores and low MSE values for both training and testing datasets. This indicates that the model effectively captures the relationship between the predictors and the performance index.

Quantile regression models provide additional insights by estimating the conditional median (0.5 quantile) and other quantiles (0.25 and 0.75) of the response variable distribution. This allows for a more nuanced understanding of how predictor variables affect different parts of the distribution of the performance index.

**Performance at Different Quantiles:**

The 0.25 quantile model shows higher MSE values, suggesting that predicting lower quantiles of the performance index is more challenging. This could be due to a higher variability in the data at these quantiles or the presence of outliers affecting the lower end.

The 0.5 quantile model performs similarly to the traditional model, indicating that it effectively captures the central tendency of the performance index.

The 0.75 quantile model also shows higher MSE values, indicating variability in predictions at the upper quantiles. This suggests that the upper quantile predictions might be influenced by factors not fully captured by the model.

**Residual Analysis:**

Residual plots and histograms reveal the presence of heteroscedasticity in the data, meaning that the variance of residuals varies across levels of the predicted values. Quantile regression helps address this by providing separate predictions for different parts of the distribution, which can offer a more comprehensive view of the data's variability.