Evaluating Factors Influencing PISA Test Reading Scores

Problem Statement:

The PISA (Program for International Student Assessment) dataset provides insight into the educational performance of 15-year-olds worldwide. This project aims to analyze various factors influencing students' reading scores. By evaluating demographics, parental education, school characteristics, and student behavior, we will identify significant predictors of reading performance. The findings can help educators and policymakers understand how to improve student outcomes.

Dataset: PISA

Dataset Overview:

The dataset includes multiple features related to students' backgrounds and their performance in reading assessments. Below are the key variables in the dataset:

Feature	Description
grade	The grade level of the student (most 15-year-olds in America are in 10th grade).
male	Indicator of whether the student is male (1 for male, 0 for female).
raceeth	Composite score representing the race/ethnicity of the student.
preschool	Indicator of whether the student attended preschool (1 for yes, 0 for no).
expectBachelors	Indicator of whether the student expects to obtain a bachelor's degree (1 for yes, 0 for no).
motherHS	Indicator of whether the student's mother completed high school (1 for yes, 0 for no).
motherBachelors	Indicator of whether the student's mother obtained a bachelor's degree (1 for yes, 0 for no).
motherWork	Indicator of whether the student's mother has part-time or full-time work (1 for yes, 0 for no).
fatherHS	Indicator of whether the student's father completed high school (1 for yes, 0 for no).

Feature	Description
fatherBachelors	Indicator of whether the student's father obtained a bachelor's degree (1 for yes, 0 for no).
fatherWork	Indicator of whether the student's father has part-time or full-time work (1 for yes, 0 for no).
selfBornUS	Indicator of whether the student was born in the United States (1 for yes, 0 for no).
motherBornUS	Indicator of whether the student's mother was born in the United States (1 for yes, 0 for no).
fatherBornUS	Indicator of whether the student's father was born in the United States (1 for yes, 0 for no).
englishAtHome	Indicator of whether the student speaks English at home (1 for yes, 0 for no).
computerForSchoolwork	Indicator of whether the student has access to a computer for schoolwork (1 for yes, 0 for no).
read30MinsADay	Indicator of whether the student reads for pleasure for 30 minutes/day (1 for yes, 0 for no).
minutesPerWeekEnglish	The number of minutes per week the student spends in English class.
studentsInEnglish	The number of students in this student's English class.
schoolHasLibrary	Indicator of whether the student's school has a library (1 for yes, 0 for no).
publicSchool	Indicator of whether the student attends a public school (1 for yes, 0 for no).
urban	Indicator of whether the student's school is in an urban area (1 for yes, 0 for no).
schoolSize	The number of students in the student's school.
readingScore	The student's reading score, on a 1000-point scale.

Project Steps:

1. Data Loading and Preprocessing

- Load the dataset and inspect the data for missing values and data types.

- Clean the data by handling missing values and converting categorical variables as needed.

2. Descriptive Statistics

- Calculate and analyze descriptive statistics for key variables, such as reading scores and study habits.

3. Data Visualization

- Create visualizations to explore relationships in the data, including box plots and scatter plots.

4. Hypothesis Testing

- Formulate hypotheses regarding the impact of specific factors (e.g., gender, parental education) on reading scores.
- Conduct appropriate statistical tests, such as t-tests or ANOVA, to evaluate these hypotheses.

5. Regression Analysis

- Use multiple regression analysis to identify significant predictors of reading scores.
- Interpret the coefficients to understand the influence of each factor.

6. Confidence Intervals

- Calculate confidence intervals for the regression coefficients to assess the reliability of the estimates.

7. Conclusion and Recommendations

- Summarize the key findings from the analysis.
- Discuss implications for educational practices and policies aimed at improving student performance.

Deliverables:

- A Python script or Jupyter Notebook containing the code for data analysis.
- A PPT report summarizing the findings, visualizations, and interpretations.