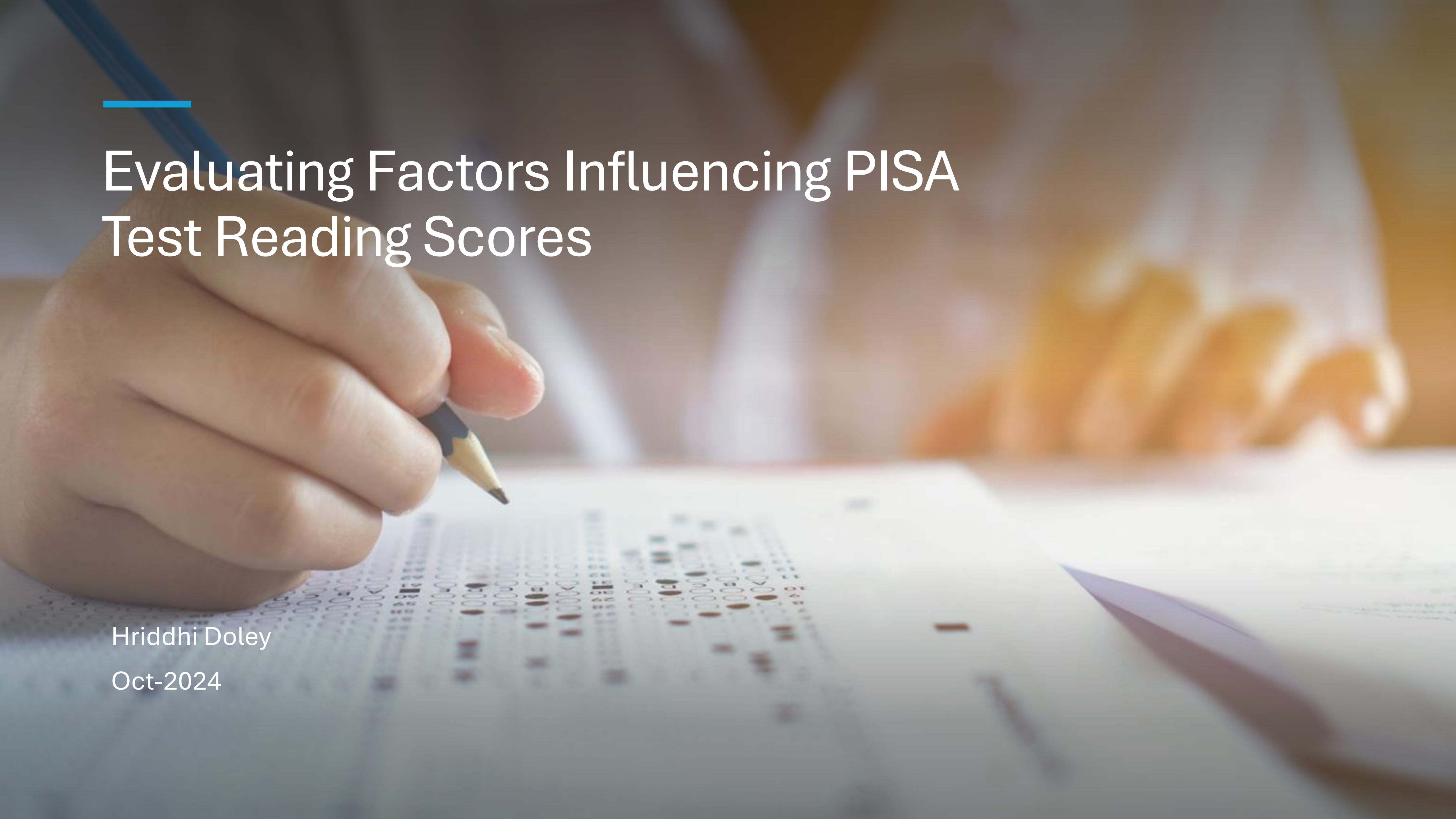


Evaluating Factors Influencing PISA Test Reading Scores

Hriddhi Doley

Oct-2024

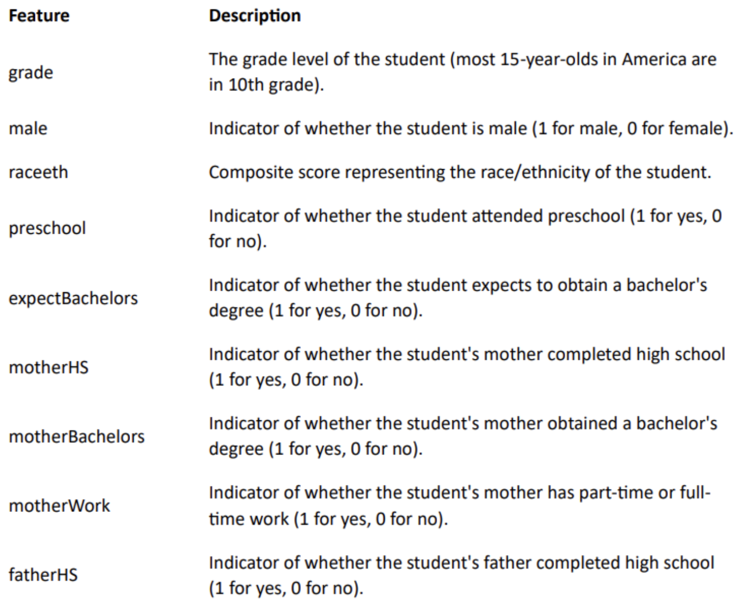
Problem Statement

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. We will identify significant predictors of reading performance by evaluating demographics, parental education, school characteristics, and student behavior. The findings can help educators and policymakers understand how to improve student outcomes.

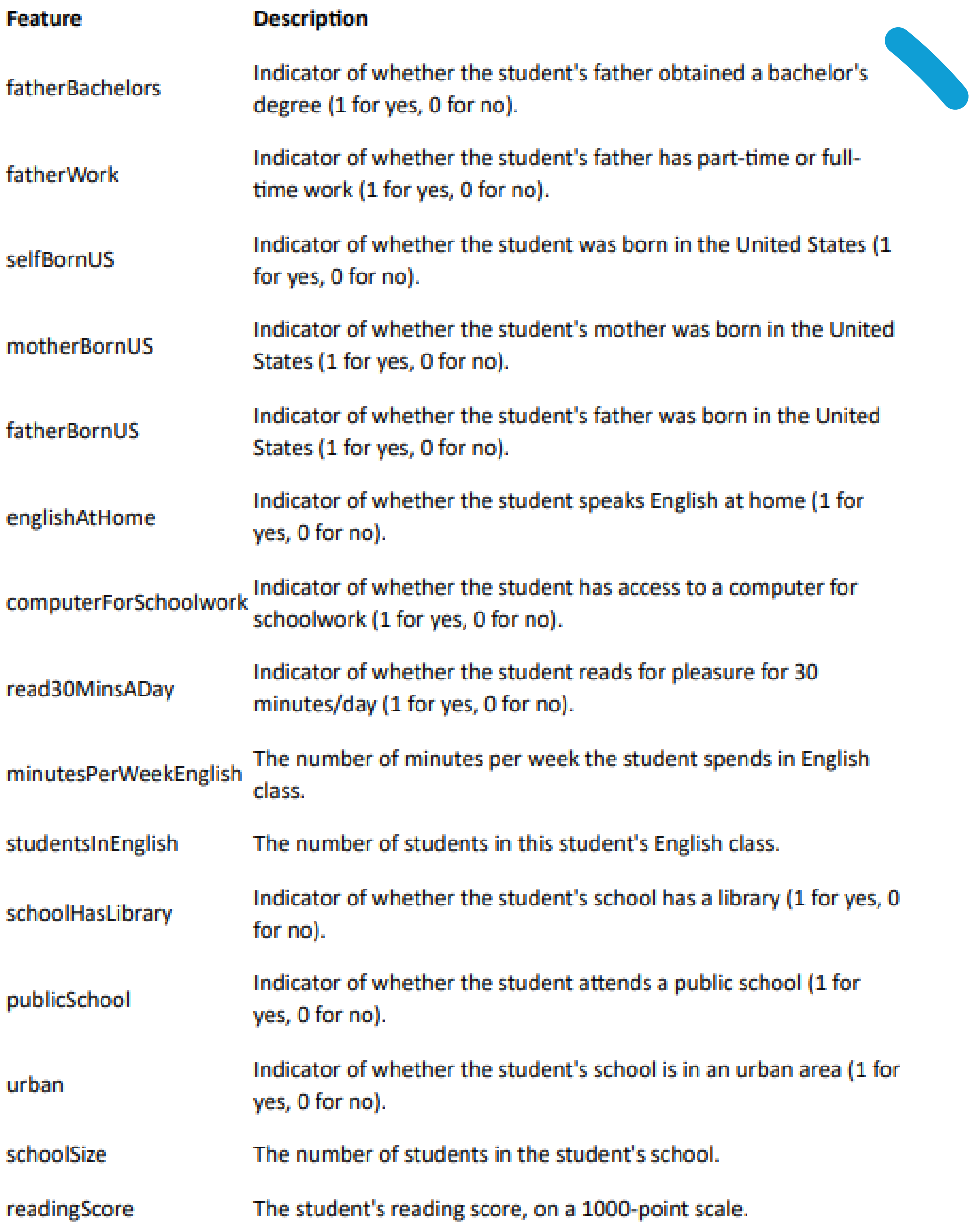
- Dataset: PISA

Dataset overview

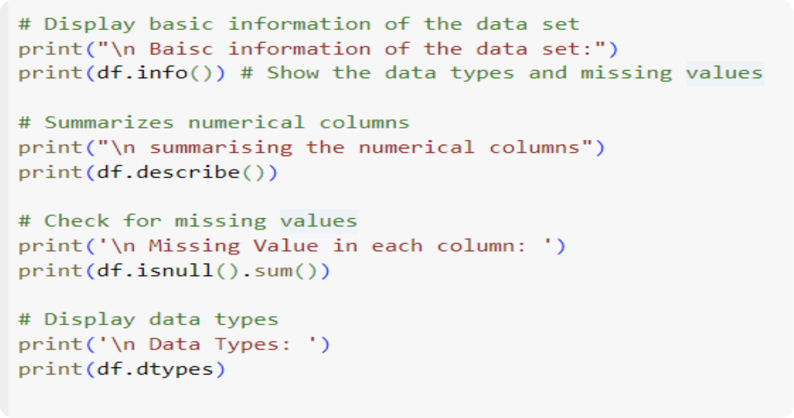
- The dataset includes multiple features related to students' backgrounds and their performance in reading assessments. Below are the key variables in the dataset:



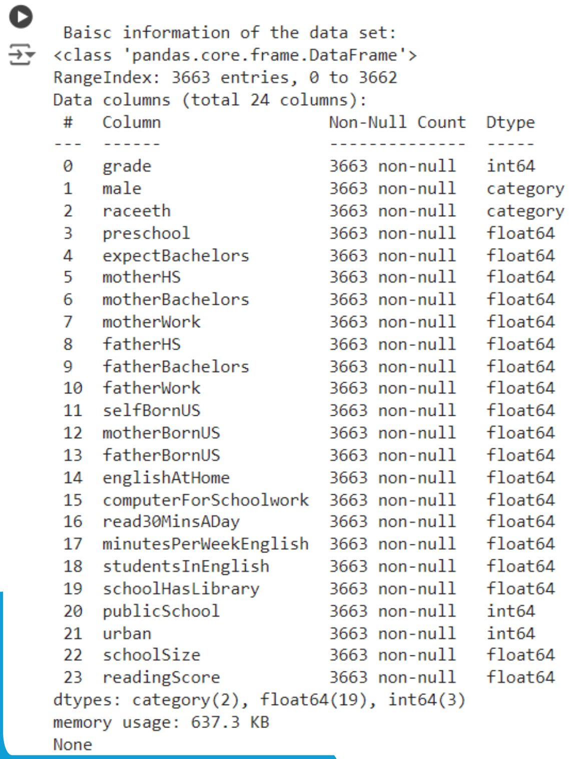
Feature	Description
grade	The grade level of the student (most 15-year-olds in America are in 10th grade).
male	Indicator of whether the student is male (1 for male, 0 for female).
raceeth	Composite score representing the race/ethnicity of the student.
preschool	Indicator of whether the student attended 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.



```
# Display basic information of the data set
print("\n Baisc information of the data set:")
print(df.info()) # Show the data types and missing values

# Summarizes numerical columns
print("\n summarising the numerical columns")
print(df.describe())

# Check for missing values
print('\n Missing Value in each column: ')
print(df.isnull().sum())

# Display data types
print('\n Data Types: ')
print(df.dtypes)
```

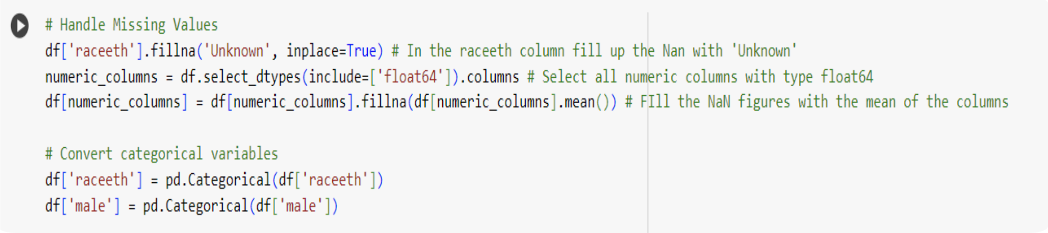


```
 Baisc information of the data set:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3663 entries, 0 to 3662
Data columns (total 24 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   grade                  3663 non-null   int64
 1   male                   3663 non-null   category
 2   raceeth                3663 non-null   category
 3   preschool              3663 non-null   float64
 4   expectBachelors        3663 non-null   float64
 5   motherHS               3663 non-null   float64
 6   motherBachelors        3663 non-null   float64
 7   motherWork             3663 non-null   float64
 8   fatherHS               3663 non-null   float64
 9   fatherBachelors        3663 non-null   float64
 10  fatherWork             3663 non-null   float64
 11  selfBornUS             3663 non-null   float64
 12  motherBornUS           3663 non-null   float64
 13  fatherBornUS           3663 non-null   float64
 14  englishAtHome          3663 non-null   float64
 15  computerForSchoolwork  3663 non-null   float64
 16  read30MinsADay         3663 non-null   float64
 17  minutesPerWeekEnglish  3663 non-null   float64
 18  studentsInEnglish      3663 non-null   float64
 19  schoolHasLibrary       3663 non-null   float64
 20  publicSchool           3663 non-null   int64
 21  urban                  3663 non-null   int64
 22  schoolSize             3663 non-null   float64
 23  readingScore           3663 non-null   float64
dtypes: category(2), float64(19), int64(3)
memory usage: 637.3 KB
None
```

Data inspection

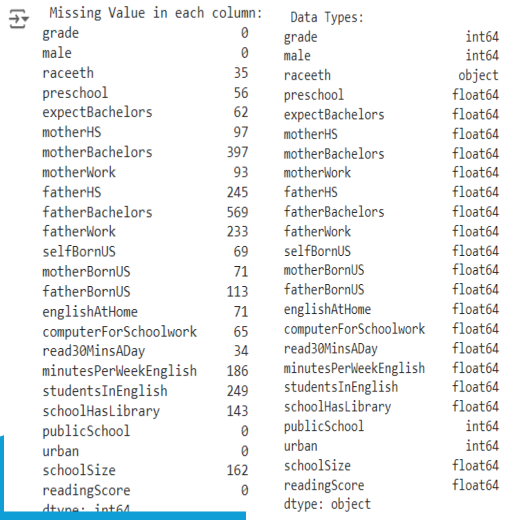
- Data types of the columns and the missing values were checked using the pandas python libraries. Below are the observations:
-
- The dataset contains 3663 entries and 24 columns.
- Most columns are of type float64, with 'grade' and 'male' being int64, and 'raceeth' being object (likely categorical).
- There are missing values in several columns, with 'fatherBachelors' having the most (569 missing values).

Handling Missing Values



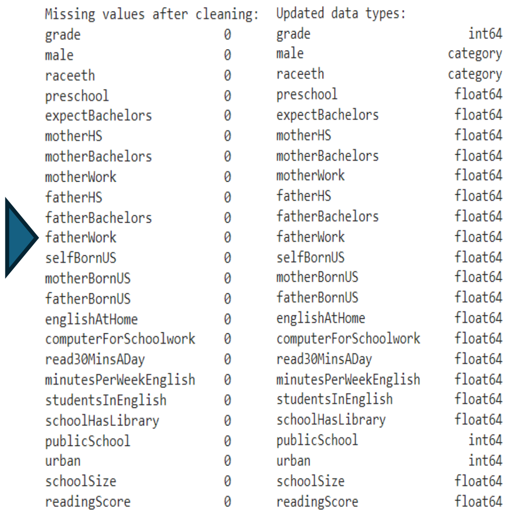
```
# Handle Missing Values
df['raceeth'].fillna('Unknown', inplace=True) # In the raceeth column fill up the Nan with 'Unknown'
numeric_columns = df.select_dtypes(include=['float64']).columns # Select all numeric columns with type float64
df[numeric_columns] = df[numeric_columns].fillna(df[numeric_columns].mean()) # FIll the NaN figures with the mean of the columns

# Convert categorical variables
df['raceeth'] = pd.Categorical(df['raceeth'])
df['male'] = pd.Categorical(df['male'])
```

Before Cleaning



```
Missing Value in each column:       Data Types:
grade                      0        grade                      int64
male                       0        male                       int64
raceeth                   35        raceeth                   object
preschool                 56        preschool                float64
expectBachelors           62        expectBachelors          float64
motherHS                  97        motherHS                 float64
motherBachelors          397        motherBachelors          float64
motherWork                93        motherWork               float64
fatherHS                 245        fatherHS                 float64
fatherBachelors          569        fatherBachelors          float64
fatherWork               233        fatherWork               float64
selfBornUS                69        selfBornUS               float64
motherBornUS              71        motherBornUS             float64
fatherBornUS             113        fatherBornUS             float64
englishAtHome             71        englishAtHome            float64
computerForSchoolwork     65        computerForSchoolwork    float64
read30MinsADay            34        read30MinsADay           float64
minutesPerWeekEnglish    186        minutesPerWeekEnglish    float64
studentsInEnglish        249        studentsInEnglish        float64
schoolHasLibrary         143        schoolHasLibrary         float64
publicSchool               0        publicSchool               int64
urban                      0        urban                      int64
schoolSize               162        schoolSize               float64
readingScore               0        readingScore             float64
dtype: int64                        dtype: object
```

After Cleaning



```
Missing values after cleaning:   Updated data types:
grade                    0       grade                    int64
male                     0       male                  category
raceeth                  0       raceeth               category
preschool                0       preschool              float64
expectBachelors          0       expectBachelors        float64
motherHS                 0       motherHS               float64
motherBachelors          0       motherBachelors        float64
motherWork               0       motherWork             float64
fatherHS                 0       fatherHS               float64
fatherBachelors          0       fatherBachelors        float64
fatherWork               0       fatherWork             float64
selfBornUS               0       selfBornUS             float64
motherBornUS             0       motherBornUS           float64
fatherBornUS             0       fatherBornUS           float64
englishAtHome            0       englishAtHome          float64
computerForSchoolwork    0       computerForSchoolwork  float64
read30MinsADay           0       read30MinsADay         float64
minutesPerWeekEnglish    0       minutesPerWeekEnglish  float64
studentsInEnglish        0       studentsInEnglish      float64
schoolHasLibrary         0       schoolHasLibrary       float64
publicSchool             0       publicSchool             int64
urban                    0       urban                    int64
schoolSize               0       schoolSize             float64
readingScore             0       readingScore           float64
```

- The following steps were followed
- Fill up the missing values in the 'raceeth' column with 'Unknown’
- In numeric columns with their mean
- Convert specified columns to categorical data types
- Verify the changes by printing missing values, data types, and summary statistics.

Data Visualization

The following three data visualizations were used to analyze the data

1.Box Plot of Reading Scores by Race/Ethnicity:
This plot shows the distribution of reading scores across different racial/ethnic groups.

2. Scatter Plot of Reading Scores vs. Minutes per Week Spent on English:
This plot illustrates the relationship between time spent on English and reading scores.

3. Box Plot of Reading Scores by Gender:
This plot compares reading scores between genders.

Key insights include:

1. The mean reading score is approximately 498.
2. Asian students have the highest mean reading score, while American Indian/Alaska Native students have the lowest.
3. There is a weak positive correlation (0.064) between minutes spent on English and reading scores.
4. Female students have a higher mean reading score compared to male students.

Data Visualization: Box Plot of Reading Scores by Race/Ethnicity

1.Distribution of Scores:

1. The box plot shows the distribution of reading scores for different racial/ethnic groups.
2. Each box represents the interquartile range (IQR) for that group, with the median shown as a line within the box.

2.Median Scores:

1. Asian students appear to have the highest median reading score, followed by White students.
2. Hispanic and Black students have lower median scores compared to Asian and White students.

3.Score Ranges:

1. There is considerable overlap in the score ranges across all groups, indicating that while there are differences in average performance, there is also significant variation within each group.
2. Asian students show the largest range of scores, suggesting high variability in performance within this group.

4.Outliers:

1. All groups have outliers, represented by individual points beyond the whiskers.
2. There are more low outliers than high outliers across all groups, indicating some students in each category who significantly underperform compared to their peers.

5.Performance Gaps:

1. There appears to be a noticeable gap in performance between Asian/White students and Hispanic/Black students.
2. This gap suggests potential disparities in educational outcomes that may need to be addressed.

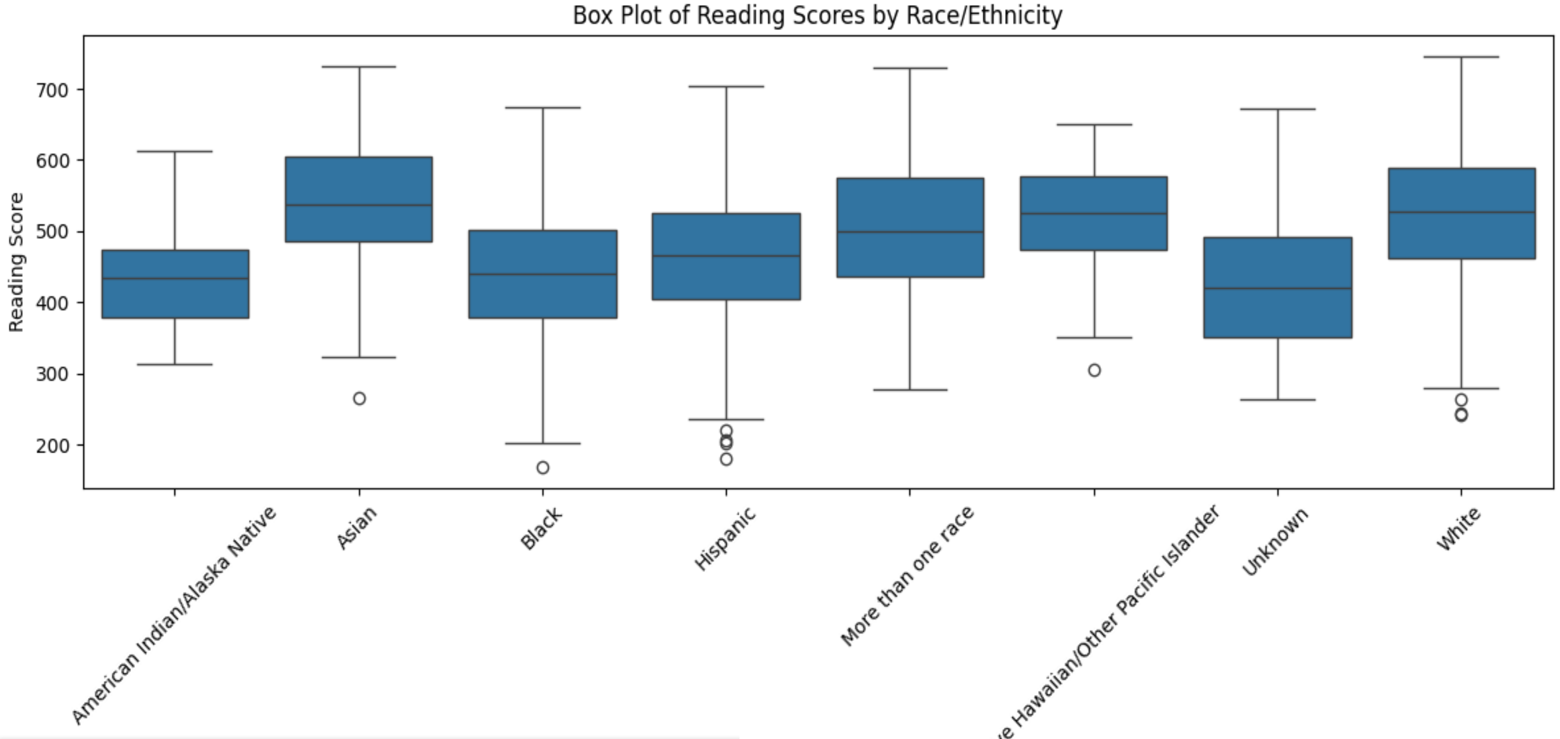
6.Variability:

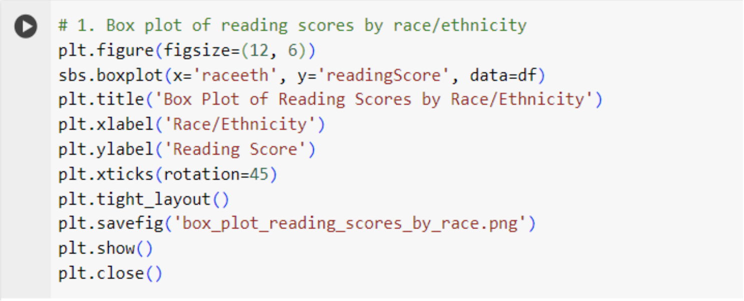
1. The size of the boxes (IQR) varies between groups, with some showing more consistency in scores (smaller boxes) and others showing more variability (larger boxes).

7.Skewness:

1. Some distributions appear slightly skewed, as indicated by the position of the median line within the box.

Insights: These insights highlight performance differences among racial/ethnic groups in reading scores. While these differences exist, many factors beyond race/ethnicity can influence academic performance, such as socioeconomic status, school quality, and individual circumstances. This data could be used to inform targeted interventions and support programs to help close achievement gaps and ensure equitable educational outcomes for all students.





```
# 1. Box plot of reading scores by race/ethnicity
plt.figure(figsize=(12, 6))
sbs.boxplot(x='raceeth', y='readingScore', data=df)
plt.title('Box Plot of Reading Scores by Race/Ethnicity')
plt.xlabel('Race/Ethnicity')
plt.ylabel('Reading Score')
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('box_plot_reading_scores_by_race.png')
plt.show()
plt.close()
```

Data Visualization: Scatter Plot of Reading Scores vs. Mins/ Week Spent on English

1.Overall Relationship: The scatter plot shows how reading scores relate to the time students spend on English each week. The correlation coefficient is about 0.0678, which means there's a very weak positive relationship between these two factors. In simple terms, spending more time on English doesn't necessarily lead to much higher reading scores.

2.Spread of Data: The dots on the graph are quite spread out, forming a cloud-like shape. This scatter suggests that while some students who spend more time on English do have higher scores, many others don't. It's not a clear-cut "more time = better scores" situation.

3.Time Spent on English:

1. On average, students spend about 266 minutes (roughly 4.5 hours) per week on English.
2. The least amount of time spent is 0 minutes, while the most is 2400 minutes (40 hours) per week.
3. Half of the students spend between 225 and 300 minutes (3.75 to 5 hours) per week on English.

4.Reading Scores:

1. The average reading score is about 498 out of 746.
2. Scores range from as low as 169 to as high as 746.
3. Half of the students score between 432 and 566.

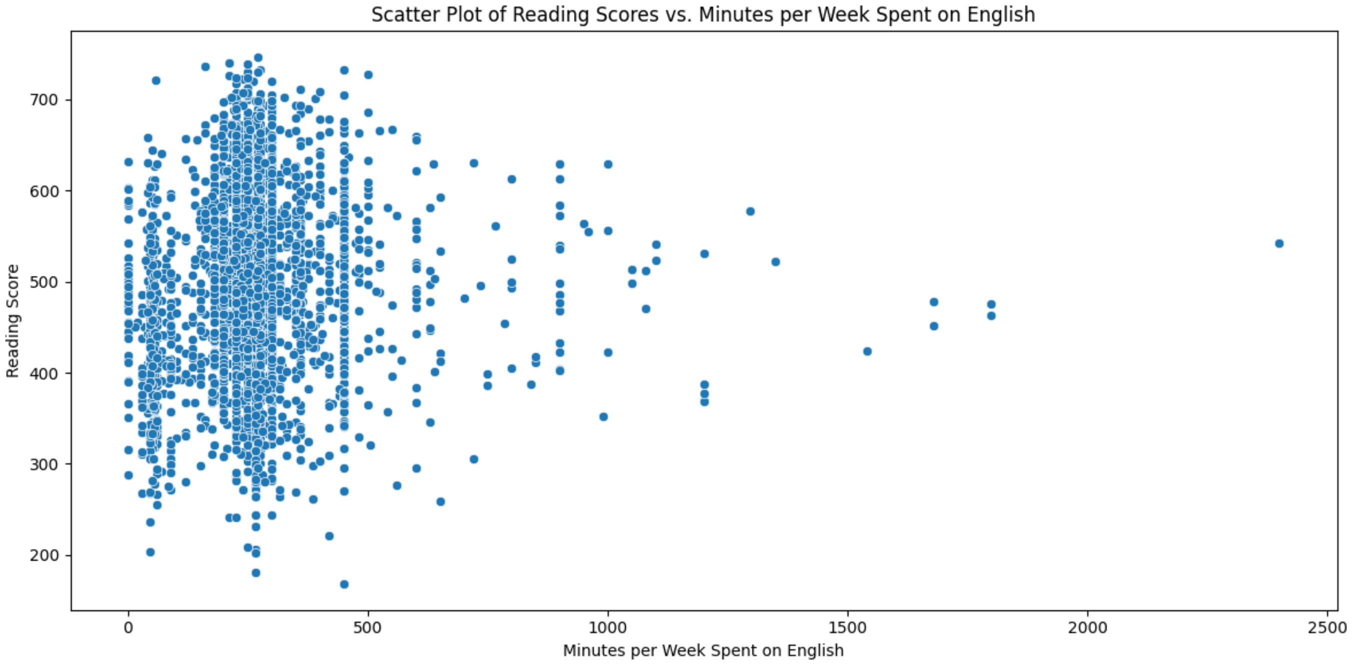
5.Interesting Points:

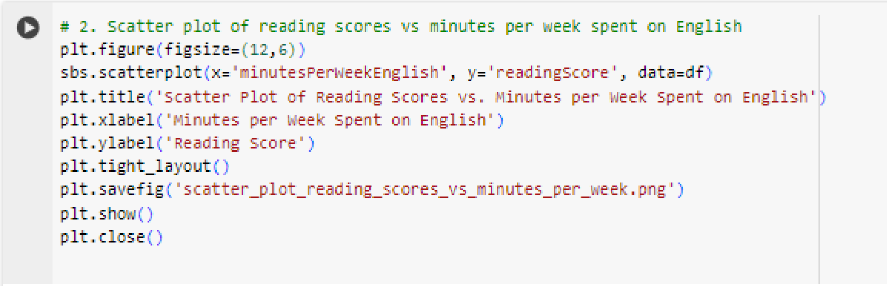
1. There are some students who spend very little time on English but still achieve high scores.
2. Conversely, some students spend a lot of time but don't necessarily score higher.
3. Most students cluster around the middle range for both time spent and scores achieved.

6.Considerations:

1. The data doesn't tell us about the quality of the time spent or the specific activities done during English study.
2. Individual differences, like learning speed or prior knowledge, aren't captured in this simple relationship.

Insights: In conclusion, while there's a slight tendency for more time spent on English to relate to higher reading scores, it's a very weak connection. This suggests that other factors beyond just time spent are important in determining reading performance. For students and educators, this means focusing on effective study strategies might be more beneficial than simply increasing study time.





```
# 2. Scatter plot of reading scores vs minutes per week spent on English
plt.figure(figsize=(12,6))
sbs.scatterplot(x='minutesPerWeekEnglish', y='readingScore', data=df)
plt.title('Scatter Plot of Reading Scores vs. Minutes per Week Spent on English')
plt.xlabel('Minutes per Week Spent on English')
plt.ylabel('Reading Score')
plt.tight_layout()
plt.savefig('scatter_plot_reading_scores_vs_minutes_per_week.png')
plt.show()
plt.close()
```

Data Visualization: Box Plot of Reading Scores by Gender

The box plot shows the distribution of reading scores for male and female students. Let's break down the insights:

1. Median Scores:

1. Female students generally have a higher median reading score compared to male students. The median is the middle value, so this suggests that, on average, females tend to score higher in reading.

2. Score Range:

1. Both genders have a similar range of scores, but the spread (or variability) of scores is slightly wider for males. This means that while some male students score very high, others score quite low, leading to more variability.

3. Interquartile Range (IQR):

1. The IQR, which is the range between the 25th and 75th percentiles, is slightly larger for males. This indicates more variation in the middle 50% of male scores compared to females.

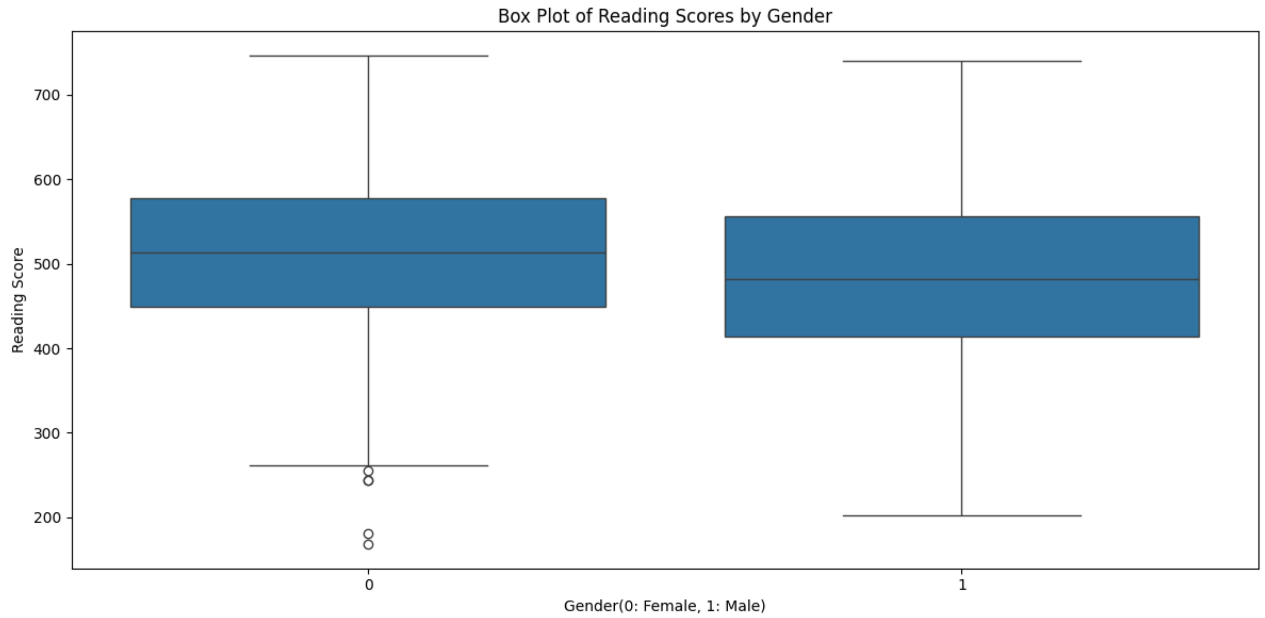
4. Outliers:

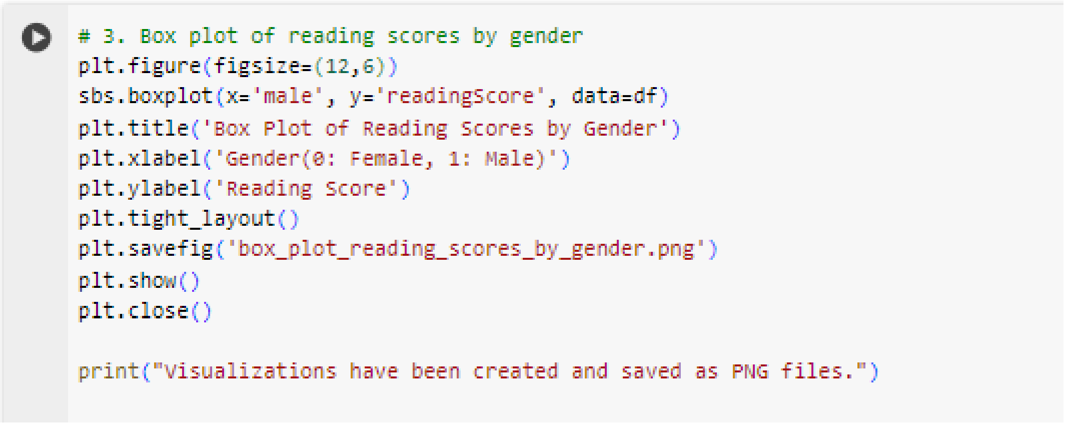
1. Both genders have outliers, which are scores that are significantly higher or lower than the rest. These are represented by individual points outside the whiskers of the box plot.

5. Overall Performance:

1. The plot suggests that female students generally perform better in reading than male students, as indicated by the higher median and slightly more compact score distribution.

In summary, the box plot indicates that female students tend to have higher reading scores on average, with less variability compared to male students. This suggests that gender may play a role in reading performance, with females generally outperforming males in this dataset.



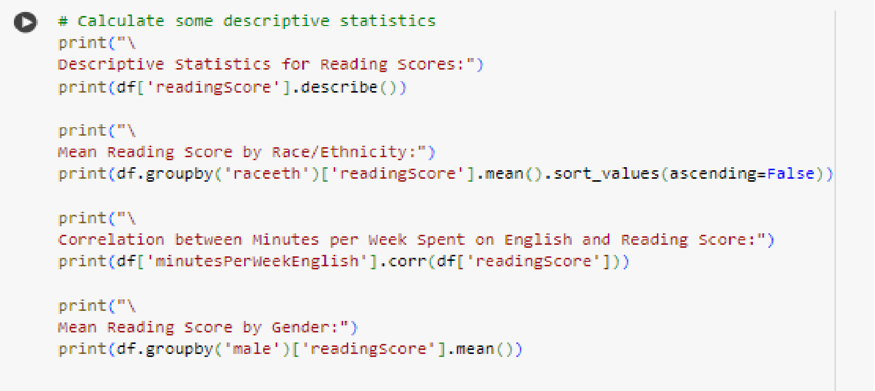


```
# 3. Box plot of reading scores by gender
plt.figure(figsize=(12,6))
sbs.boxplot(x='male', y='readingScore', data=df)
plt.title('Box Plot of Reading Scores by Gender')
plt.xlabel('Gender(0: Female, 1: Male)')
plt.ylabel('Reading Score')
plt.tight_layout()
plt.savefig('box_plot_reading_scores_by_gender.png')
plt.show()
plt.close()

print("Visualizations have been created and saved as PNG files.")
```

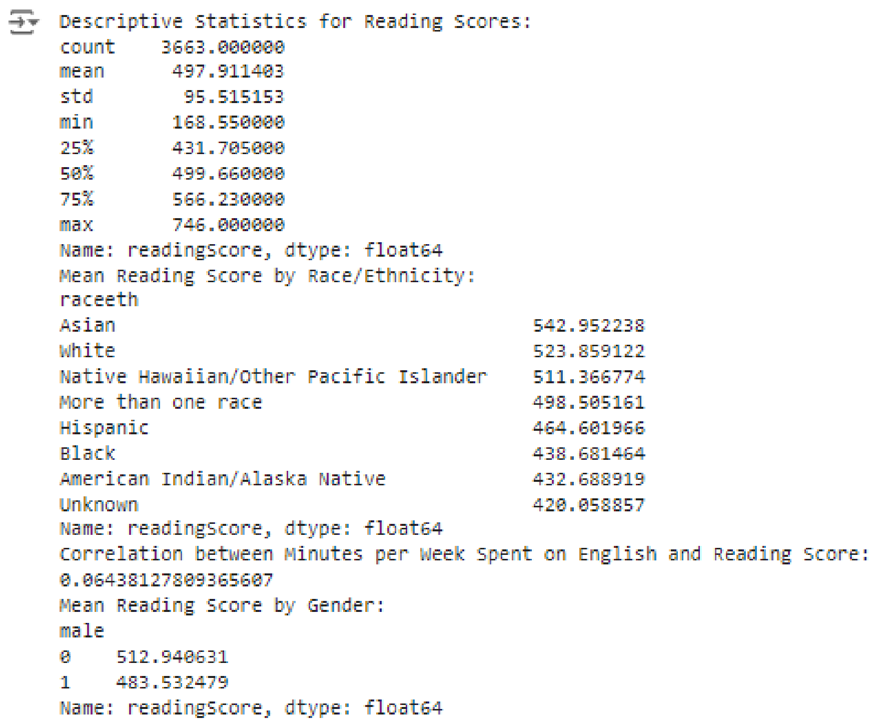
Descriptive Statistics - Insights

The below insights highlight differences in reading performance based on gender and race/ethnicity, while also showing that time spent on English has a minimal impact on reading scores.

1. **Descriptive Statistics for Reading Scores**:
   - The average reading score is about 498.
   - Scores range from a low of about 169 to a high of 746.
   - Most students score between 432 and 566, with the middle score (median) being around 500.
2. **Mean Reading Score by Race/Ethnicity**:
   - Asian students have the highest average reading score, followed by White students.
   - American Indian/Alaska Native and Black students have the lowest average scores.
   - This suggests some differences in reading performance across different racial/ethnic groups.
3. **Correlation between Minutes per Week Spent on English and Reading Score**:
   - The correlation is about 0.068, which is very weak.
   - This means that spending more time on English doesn't strongly relate to higher reading scores.
4. **Mean Reading Score by Gender**:
   - Female students (coded as 0) have a higher average reading score (about 513) compared to male students (coded as 1) who have an average score of about 484.
   - This indicates that, on average, female students perform better in reading than male students.



```
# Calculate some descriptive statistics
print("\
Descriptive Statistics for Reading Scores:")
print(df['readingScore'].describe())

print("\
Mean Reading Score by Race/Ethnicity:")
print(df.groupby('raceeth')['readingScore'].mean().sort_values(ascending=False))

print("\
Correlation between Minutes per Week Spent on English and Reading Score:")
print(df['minutesPerWeekEnglish'].corr(df['readingScore']))

print("\
Mean Reading Score by Gender:")
print(df.groupby('male')['readingScore'].mean())
```

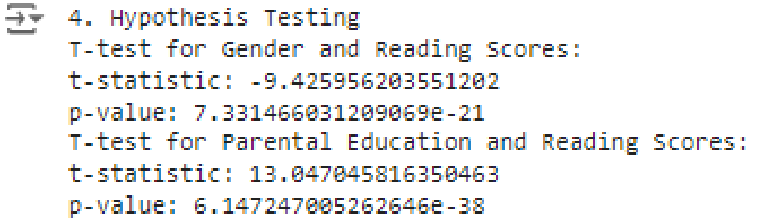


```
Descriptive Statistics for Reading Scores:
count    3663.000000
mean      497.911403
std        95.515153
min       168.550000
25%       431.705000
50%       499.660000
75%       566.230000
max       746.000000
Name: readingScore, dtype: float64
Mean Reading Score by Race/Ethnicity:
raceeth
Asian                                       542.952238
White                                       523.859122
Native Hawaiian/Other Pacific Islander      511.366774
More than one race                          498.505161
Hispanic                                    464.601966
Black                                       438.681464
American Indian/Alaska Native               432.688919
Unknown                                     420.058857
Name: readingScore, dtype: float64
Correlation between Minutes per Week Spent on English and Reading Score:
0.06438127809365607
Mean Reading Score by Gender:
male
0    512.940631
1    483.532479
Name: readingScore, dtype: float64
```

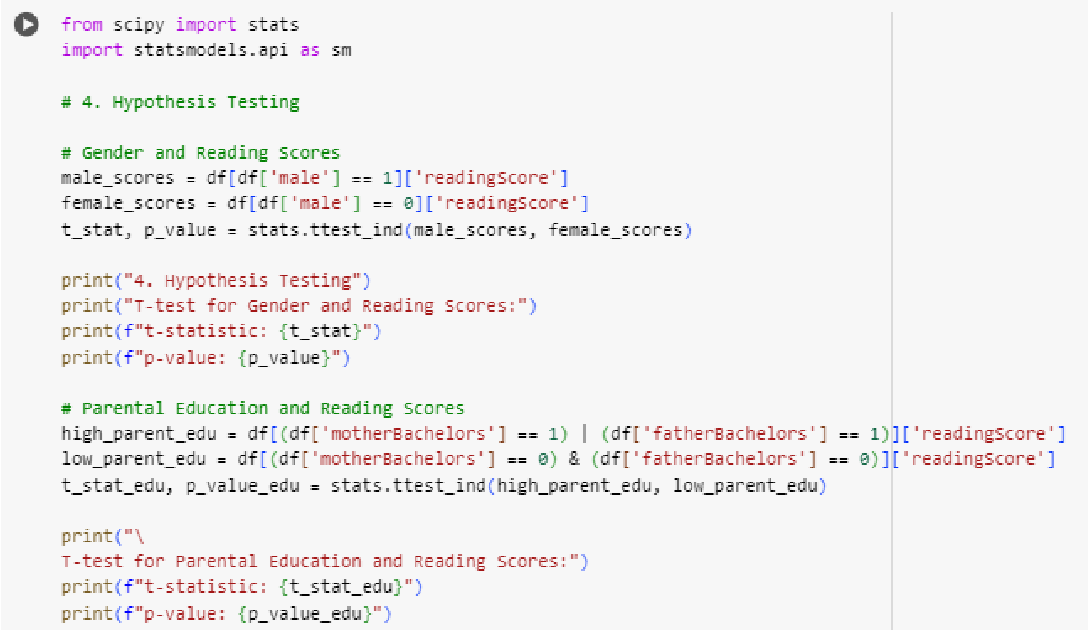
Hypothesis Testing

1. **Gender and Reading Scores:**
The t-test for gender shows a significant difference in reading scores between males and females with a t-statistic of -9.43 and a p-value of 7.33×10–21, indicating that gender has a significant impact on reading scores. A highly negative t-statistic suggests that the average reading scores for males are significantly lower than the females. The p-value is extremely small, meaning this difference is statistically significant, so it's highly unlikely this result is due to chance.

2. **Parental Education and Reading Scores:**
The t-test for parental education shows a significant difference in reading scores between students with at least one parent having a bachelor's degree and those without, with a t-statistic of 13.05 and a p-value of 6.15×10–38. The positive t-statistic suggests that there is a strong association between higher parental education levels and better reading scores. The extremely small p-value again means this result is statistically significant, implying that it's very unlikely this association is due to random variation.



```
4. Hypothesis Testing
T-test for Gender and Reading Scores:
t-statistic: -9.425956203551202
p-value: 7.331466031209069e-21
T-test for Parental Education and Reading Scores:
t-statistic: 13.047045816350463
p-value: 6.147247005262646e-38
```



```python
from scipy import stats
import statsmodels.api as sm

# 4. Hypothesis Testing

# Gender and Reading Scores
male_scores = df[df['male'] == 1]['readingScore']
female_scores = df[df['male'] == 0]['readingScore']
t_stat, p_value = stats.ttest_ind(male_scores, female_scores)

print("4. Hypothesis Testing")
print("T-test for Gender and Reading Scores:")
print(f"t-statistic: {t_stat}")
print(f"p-value: {p_value}")

# Parental Education and Reading Scores
high_parent_edu = df[(df['motherBachelors'] == 1) | (df['fatherBachelors'] == 1)]['readingScore']
low_parent_edu = df[(df['motherBachelors'] == 0) & (df['fatherBachelors'] == 0)]['readingScore']
t_stat_edu, p_value_edu = stats.ttest_ind(high_parent_edu, low_parent_edu)

print("\
T-test for Parental Education and Reading Scores:")
print(f"t-statistic: {t_stat_edu}")
print(f"p-value: {p_value_edu}")
```

Regression Analysis

1. The regression model identifies significant predictors of reading scores, including gender, parental education, English spoken at home, computer use for schoolwork, and reading for 30 minutes a day.

2. R-squared (0.165): This tells us that the independent variables (factors) explain about 16.5% of the variation in reading scores. This isn't very high, suggesting many other factors influence reading scores not included in this model.

3. Adj. R-squared (0.164): Adjusted for the number of predictors, this value is close to the R-squared value, indicating the model is not overfitting.

4. Gender has a significant negative impact on reading scores, with boys scoring lower than girls.

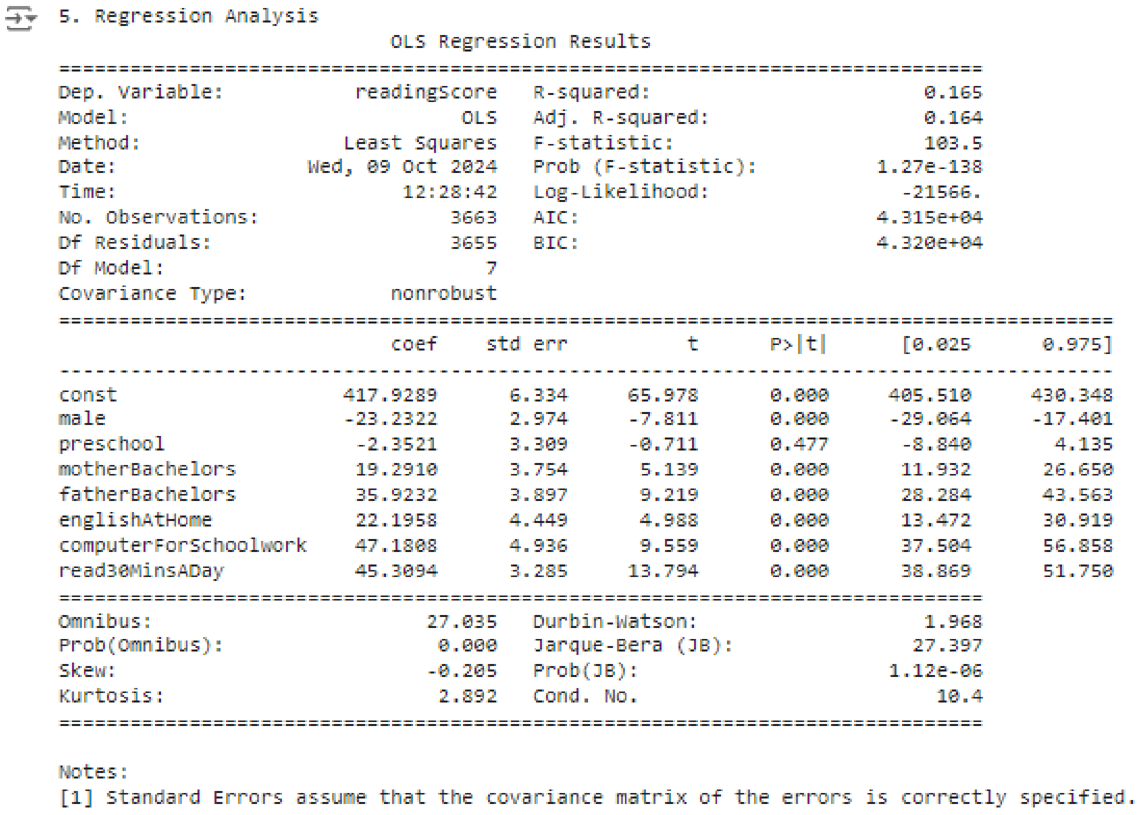
5. Parental education (both mother's and father's bachelor's degrees) has a strong positive impact on reading scores.

6. Speaking English at home, using a computer for schoolwork, and reading for 30 minutes a day are all associated with higher reading scores.

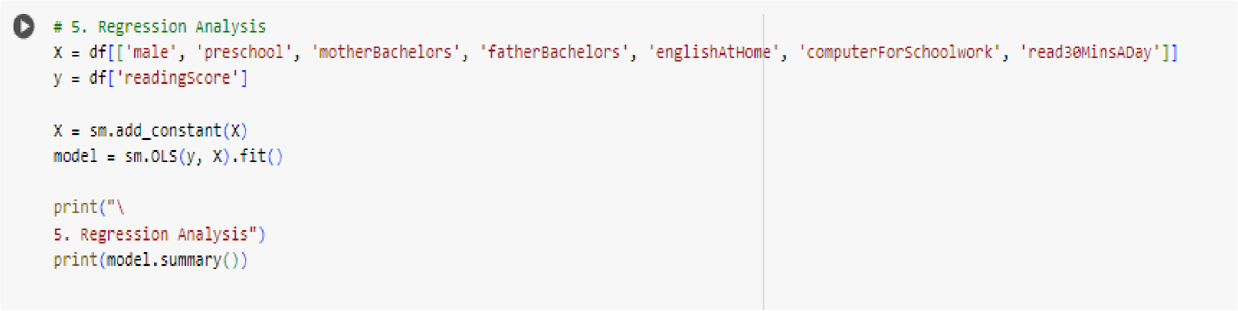
7. Attending preschool does not have a statistically significant impact on reading scores, at least in this model.

8. Some factors, like gender, parental education, and certain habits (computer use and reading), play a significant role in reading performance.

9. The overall model explains about 16.5% of the variation in reading scores, meaning that other factors not included in the analysis also influence reading performance.



```
5. Regression Analysis
                            OLS Regression Results
==============================================================================
Dep. Variable:            readingScore   R-squared:                       0.165
Model:                             OLS   Adj. R-squared:                  0.164
Method:                  Least Squares   F-statistic:                     103.5
Date:                 Wed, 09 Oct 2024   Prob (F-statistic):          1.27e-138
Time:                         12:28:42   Log-Likelihood:                 -21566.
No. Observations:                 3663   AIC:                         4.315e+04
Df Residuals:                     3655   BIC:                         4.320e+04
Df Model:                            7
Covariance Type:             nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   417.9289      6.334     65.978      0.000     405.510     430.348
male                    -23.2322      2.974     -7.811      0.000     -29.064     -17.401
preschool                -2.3521      3.309     -0.711      0.477      -8.840       4.135
motherBachelors          19.2910      3.754      5.139      0.000      11.932      26.650
fatherBachelors          35.9232      3.897      9.219      0.000      28.284      43.563
englishAtHome            22.1958      4.449      4.988      0.000      13.472      30.919
computerForSchoolwork    47.1808      4.936      9.559      0.000      37.504      56.858
read30MinsADay           45.3094      3.285     13.794      0.000      38.869      51.750
==============================================================================
Omnibus:                       27.035   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.397
Skew:                          -0.205   Prob(JB):                     1.12e-06
Kurtosis:                       2.892   Cond. No.                         10.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

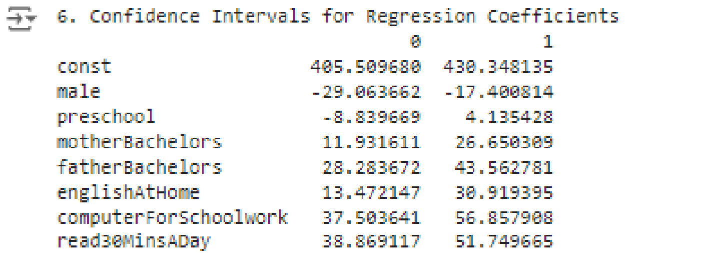


```
# 5. Regression Analysis
X = df[['male', 'preschool', 'motherBachelors', 'fatherBachelors', 'englishAtHome', 'computerForSchoolwork', 'read30MinsADay']]
y = df['readingScore']

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()

print("\
5. Regression Analysis")
print(model.summary())
```

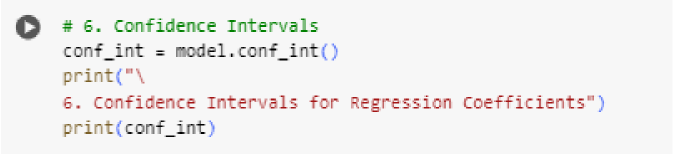
Confidence Interval

Confidence intervals for the regression coefficients provide a range of values within which the true coefficient is likely to fall, indicating the reliability of the estimates.

The intervals tell us how much each factor affects reading scores with 95% confidence. If the interval is far from zero, like for **Male**, **Father's Education**, **Computer Use**, and **Daily Reading**, we can be fairly confident these factors have a significant effect on reading scores. For **Preschool**, the confidence interval includes both negative and positive values, meaning the effect is uncertain.



```
6. Confidence Intervals for Regression Coefficients
                                0           1
const                  405.509680  430.348135
male                   -29.063662  -17.400814
preschool               -8.839669    4.135428
motherBachelors         11.931611   26.650309
fatherBachelors         28.283672   43.562781
englishAtHome           13.472147   30.919395
computerForSchoolwork   37.503641   56.857908
read30MinsADay          38.869117   51.749665
```



```
# 6. Confidence Intervals
conf_int = model.conf_int()
print("\
6. Confidence Intervals for Regression Coefficients")
print(conf_int)
```

Conclusion

Implications for Educational Practices and Policies

- **Gender Differences:** The significant impact of gender on reading scores suggests the need for targeted interventions to support male students in improving their reading skills.

- **Parental Education:** The influence of parental education highlights the importance of parental involvement and support in a child's education. Programs that engage parents and provide resources for educational support could be beneficial.

- **Predictors of Success:** Factors such as English spoken at home, computer use for schoolwork, and regular reading habits are significant predictors of reading success. Educational policies should encourage these practices to enhance student performance.

- These findings can guide educators and policymakers in developing strategies to improve reading outcomes for students.