Question2:

import pandas as pd

from sklearn.preprocessing import LabelEncoder

# Load the dataset from a CSV file

dataset = pd.read\_csv('StudentsPerformance.csv')

# Identify any missing values in the dataset

missing\_values = dataset.isnull().sum()

print("Missing Values:")

print(missing\_values)

# No missing values were found

# Display the data types of each column

print("\nData Types:")

print(dataset.dtypes)

# Normalize the column names to be lowercase and replace spaces with underscores

dataset.columns = dataset.columns.str.lower().str.replace(' ', '\_')

# Show the initial rows of the dataset

print("\nFirst few rows of the dataset:")

print(dataset.head())

# Encode the 'test preparation course' column

# 'none' will be encoded as 0 and 'completed' will be encoded as 1

label\_encoder = LabelEncoder()

dataset['test\_preparation\_course'] = label\_encoder.fit\_transform(dataset['test\_preparation\_course'])

# Save the cleaned dataset to a new CSV file

dataset.to\_csv('cleaned\_StudentPerformance.csv', index=False)

import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

# Load the cleaned dataset

data = pd.read\_csv('cleaned\_StudentPerformance.csv')

# Correlation analysis

# Calculate the correlation matrix for selected scores

correlation\_matrix = data[['math\_score', 'reading\_score', 'writing\_score']].corr()

# Compare scores for students who completed and did not complete the test preparation course using bar plots

plt.figure(figsize=(10, 6))

sns.barplot(x='test\_preparation\_course', y='math\_score', data=data, ci=None)

sns.barplot(x='test\_preparation\_course', y='reading\_score', data=data, ci=None)

sns.barplot(x='test\_preparation\_course', y='writing\_score', data=data, ci=None)

plt.title('Comparison of Mean Scores by Test Preparation Course Completion')

plt.xlabel('Test Preparation Course')

plt.ylabel('Mean Score')

plt.legend(['Math Score', 'Reading Score', 'Writing Score'])

plt.show()

# Visualize the correlation matrix as a heatmap

plt.figure(figsize=(8, 6))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", square=True)

plt.title('Correlation Matrix of Math, Reading, and Writing Scores')

plt.show()

# Stacked bar chart of scores by race/ethnicity

plt.figure(figsize=(10, 6))

sns.barplot(x='race/ethnicity', y='math\_score', data=data, ci=None)

sns.barplot(x='race/ethnicity', y='reading\_score', data=data, ci=None)

sns.barplot(x='race/ethnicity', y='writing\_score', data=data, ci=None)

plt.title('Stacked Bar Chart of Race/Ethnicity and Scores')

plt.xlabel('Race/Ethnicity')

plt.ylabel('Mean Score')

plt.legend(['Math Score', 'Reading Score', 'Writing Score'])

plt.show()

# Histogram of math scores

plt.figure(figsize=(8, 6))

sns.histplot(data['math\_score'], bins=10, kde=True)

plt.title('Histogram of Math Scores')

plt.xlabel('Math Score')

plt.ylabel('Frequency')

plt.show()

# Box plot of reading scores by gender

plt.figure(figsize=(8, 6))

sns.boxplot(x='gender', y='reading\_score', data=data)

plt.title('Box Plot of Reading Scores by Gender')

plt.xlabel('Gender')

plt.ylabel('Reading Score')

plt.show()

# Scatter plot of writing scores versus math scores

plt.figure(figsize=(8, 6))

sns.scatterplot(x='math\_score', y='writing\_score', data=data)

plt.title('Scatter Plot of Writing Scores vs Math Scores')

plt.xlabel('Math Score')

plt.ylabel('Writing Score')

plt.show()

# Bar plot of average writing score by parental level of education

plt.figure(figsize=(10, 6))

sns.barplot(x='parental\_level\_of\_education', y='writing\_score', data=data, ci=None)

plt.title('Average Writing Score by Parental Level of Education')

plt.xlabel('Parental Level of Education')

plt.ylabel('Average Writing Score')

plt.xticks(rotation=45)

plt.show()

# Pie chart showing the distribution of lunch types

lunch\_distribution = data['lunch'].value\_counts()

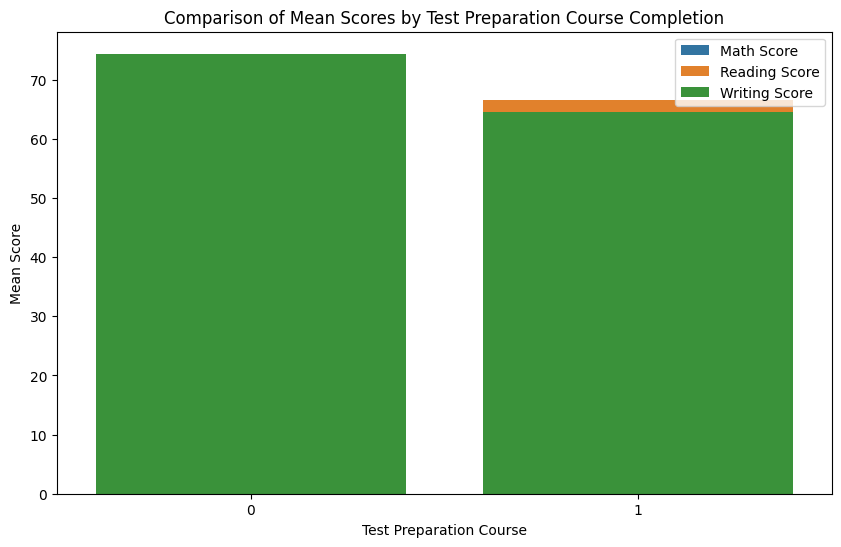
plt.figure(figsize=(8, 8))

plt.pie(lunch\_distribution, labels=lunch\_distribution.index, autopct='%1.1f%%', colors=['lightgreen', 'lightblue'])

plt.title('Distribution of Lunch Types')

plt.show()

1. Bar Plots of Mean Scores by Completion of Test-Preparation Course:



Analysis: Using this representation, it is simpler to determine whether test preparation course completion significantly affects math, reading, and writing scores. Disparities or patterns can be found fast.

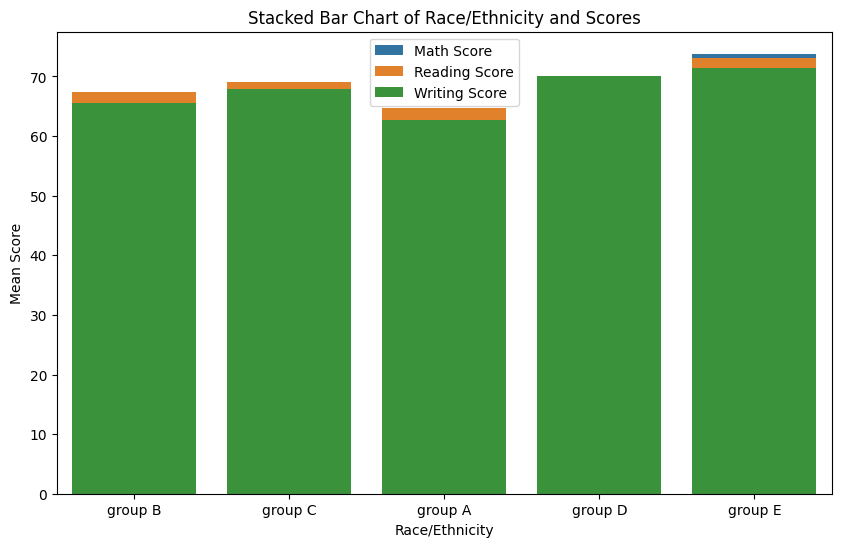
1. Correlation Matrix Heatmap:

A screenshot of a graph

Description automatically generated

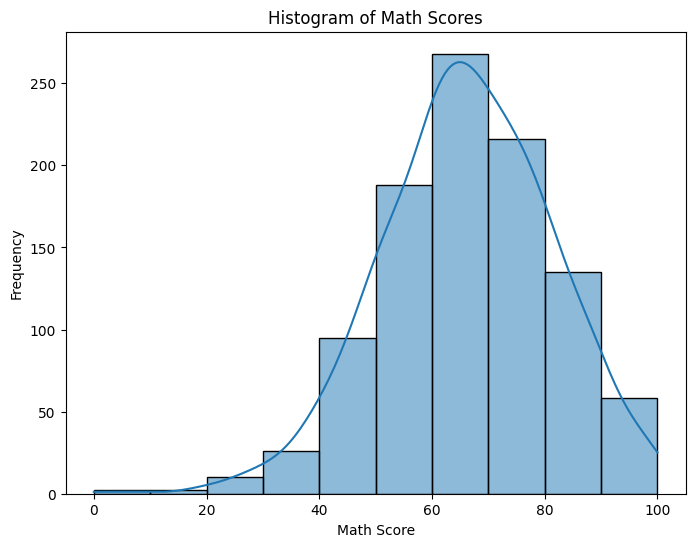
Analysis: The heatmap facilitates comprehension of the direction and degree of correlations among these scores. Strong associations are indicated by high correlation values (near to 1 or -1) which make it easier to determine which individuals often have similar performance levels.

1. Scores by Race and Ethnicity in a Stacked Bar Chart:



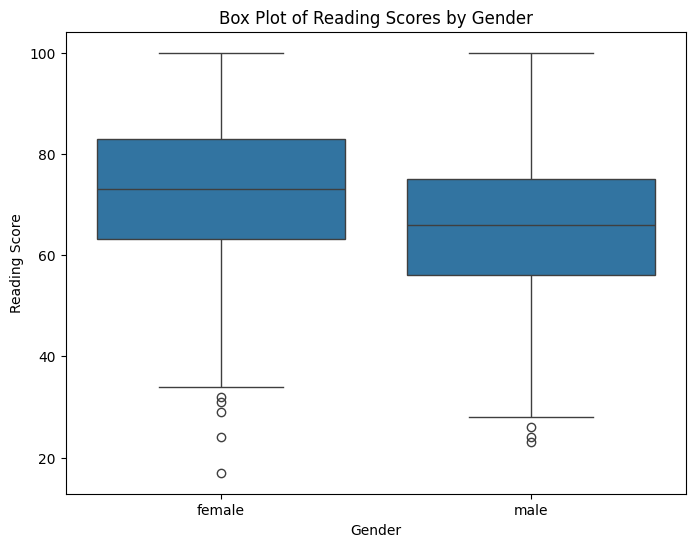
Analysis: By comparing performance across different racial/ethnic groups, this graphic aids in determining whether disparities exist. It makes it easier to identify trends and differences in educational results across demographics.

1. Histogram of Math Scores:



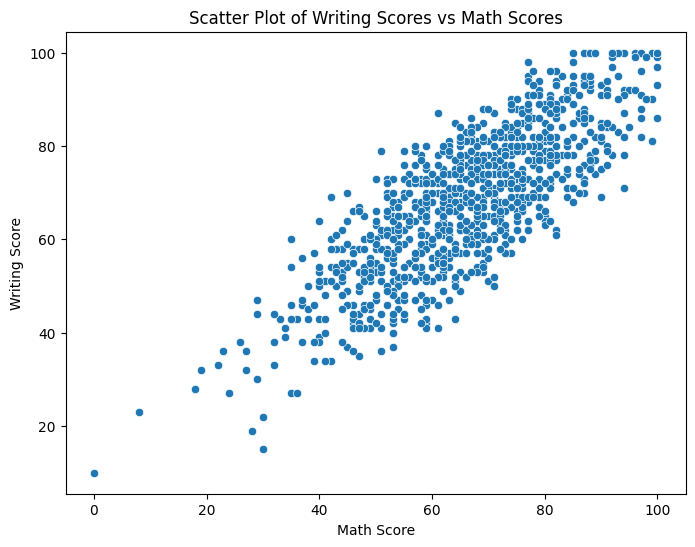
Analysis: The histogram allows for the analysis of the spread and central tendency of math scores. It helps in identifying the most common score ranges, the presence of any outliers, and the overall distribution shape (normal, skewed, etc.).

1. Box Plot of Reading Scores by Gender:



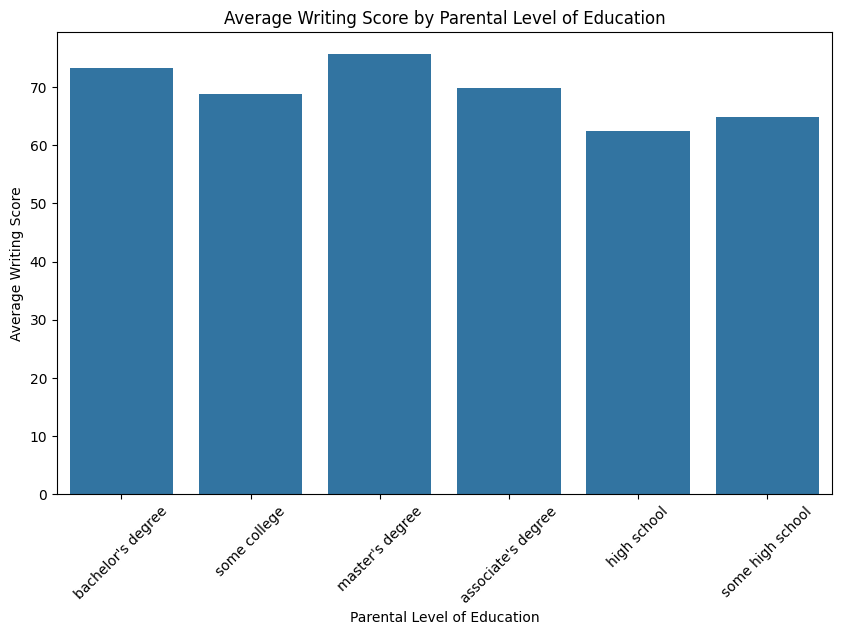
Analysis: The box plot provides insights into the median, quartiles, and potential outliers of reading scores by gender. It helps in visualizing differences in score distributions and central tendencies between male and female students.

1. Scatter Plot of Writing Scores vs. Math Scores:



By highlighting trends, clusters, and possible outliers in the data, the scatter plot allows the observation of any relationships or patterns between writing and math results.

1. Bar Plot of Average Writing Score by Parental Education Level:



This graphical representation makes it easier to understand how parental educational level affects students' writing abilities. It draws attention to any patterns or notable variations in scores according to the parents' varying educational backgrounds.