ASSIGNMENT-4

ADITYA SINGH

INTERNZLEARN TRAINING AND INTERNSHIP

NLP RELATED DATASET OPERATIONS AND WORKING WITH VARIOUS TRAINING AND TESTING ***TECHNIQUES***

In this assignment I had used several techniques for making a comprehensive analysis of the student-based performance dataset which I had imported from Kaggle. While examining the datasets of students' academic performance in mathematics and Portuguese, my aim was to uncover patterns and correlations that influence final grades. The analysis includes visualizing data distributions, studying the impact of various factors like study time and alcohol consumption on grades, and generating insights. Additionally, a linear regression model is trained to predict student outcomes, enhancing our understanding of academic performance. The project integrates heuristic-based NLP operations to provide actionable warnings, suggestions, and insights for improving student success.

VERSION OF PROGRAMMING LANGUAGE AND COMPILER

In This project, I used Visual Studio Code version 3.12.3 (64-bit) as the programming language compiler. It comes with robust features, such as IntelliSense, debugging, and integrated terminal, file path reading and provide a comprehensive environment for efficient coding and testing. Python was chosen for this analysis due to its simplicity, readability, and extensive libraries suited for data science, machine learning, and NLP tasks. Libraries like **pandas, matplotlib, seaborn, and scikit-learn** offer powerful tools for **data manipulation, visualization, and predictive modelling**, making Python the ideal choice for handling complex datasets and extracting meaningful insights in this **student performance analysis** project.

CODE IMPLEMENTATION AND RESULT PROCEDURE

Import Libraries:

- Import necessary libraries for data manipulation, visualization, and machine learning.

- **File Paths**:
  - Define file paths for the datasets.

- **Load Datasets**:
  - Load the student performance datasets using `pandas`.

- **Analyze Student Data**:
  - `analyze_student_data` function performs analysis and generates visualizations.
  - **Calculate Summary Statistics**:
    - Calculate and print summary statistics for the dataset.
  - **Correlation Matrix**:
    - Compute and plot the correlation matrix of numerical columns.
  - **Generate Insights from Correlation Matrix**:
    - Generate heuristic-based insights (warnings, suggestions, positives, negatives) based on strong correlations.
  - **Plot Distributions and Relationships**:
    - Plot various distributions and relationships such as final grades, study time, and alcohol consumption.

- **Machine Learning**:
  - Preprocess data, train a linear regression model, and make predictions.
  - **Evaluate Model**:
    - Compute and print mean squared error and R^2 score.
    - Plot true vs predicted final grades.

- **Generate Insights**:
  - **`generate_insights_from_correlation`** function extracts insights from the correlation matrix.
  - **NLP Operations**:
    - Generate warnings, suggestions, positives, and negatives using heuristic rules.
  - **Additional Insights**:
    - Generate additional insights based on specific columns (study time, alcohol consumption).
    - Analyzed both the Mathemetics and Portuguese datasets.

Machine learning algorithms and some generated insights

- **Machine Learning**:
  - Preprocess data, train a linear regression model, and make predictions.
  - **Evaluate Model**:
    - Compute and print mean squared error and R^2 score.
    - Plot true vs predicted final grades.
- **Generate Insights**:
  - `generate_insights_from_correlation` function extracts insights from the correlation matrix.
  - **NLP Operations**:
    - Generate warnings, suggestions, positives, and negatives using heuristic rules.
  - **Additional Insights**:
    - Generate additional insights based on specific columns (study time, alcohol consumption)

1. **Run Analysis**:
   - Analyzed student performance data for mathematics and Portuguese datasets.This script not only performs data analysis and visualization but also integrates machine learning for predictions and NLP-like heuristic rules to generate insights and recommendations.

The code

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Define file paths for the datasets
mat_file_path = r"C:\Users\Aditya Singh\Desktop\nlp assignment\student-mat.csv"
por_file_path = r"C:\Users\Aditya Singh\Desktop\nlp assignment\student-por.csv"
txt_file_path = r"C:\Users\Aditya Singh\Desktop\nlp assignment\student.txt"

# Load the datasets
student_mat_data = pd.read_csv(mat_file_path, sep=';')
student_por_data = pd.read_csv(por_file_path, sep=';')

# Function to perform analyses and generate visualizations
def analyze_student_data(student_data, title_suffix):
    # Calculate summary statistics
    summary_statistics = student_data.describe()
    print(f"Summary Statistics for {title_suffix}:\n", summary_statistics)

    # Select only numerical columns for correlation matrix
    numerical_columns = student_data.select_dtypes(include=['number']).columns
    correlation_matrix = student_data[numerical_columns].corr()

    # Plot the correlation matrix
    plt.figure(figsize=(14, 10))
    sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
    plt.title(f'Correlation Matrix of Student Performance Data ({title_suffix})')
    plt.show()

    # Generate insights based on correlation matrix
    generate_insights_from_correlation(correlation_matrix, title_suffix)

    # Plot the distribution of final grades (G3)
    plt.figure(figsize=(10, 6))
    sns.histplot(student_data['G3'], bins=20, kde=True)
    plt.title(f'Distribution of Final Grades (G3) ({title_suffix})')
    plt.xlabel('Final Grade (G3)')
    plt.ylabel('Frequency')
    plt.show()
```

```
    # Plot the impact of study time on final grades (G3)
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='studytime', y='G3', data=student_data)
    plt.title(f'Impact of Study Time on Final Grades (G3) ({title_suffix})')
    plt.xlabel('Study Time')
    plt.ylabel('Final Grade (G3)')
    plt.show()

    # Plot the impact of weekday alcohol consumption (Dalc) on final grades (G3)
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='Dalc', y='G3', data=student_data)
    plt.title(f'Impact of Weekday Alcohol Consumption on Final Grades (G3)
({title_suffix})')
    plt.xlabel('Weekday Alcohol Consumption (Dalc)')
    plt.ylabel('Final Grade (G3)')
    plt.show()

    # Plot the impact of weekend alcohol consumption (Walc) on final grades (G3)
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='Walc', y='G3', data=student_data)
    plt.title(f'Impact of Weekend Alcohol Consumption on Final Grades (G3)
({title_suffix})')
    plt.xlabel('Weekend Alcohol Consumption (Walc)')
    plt.ylabel('Final Grade (G3)')
    plt.show()

    # Generate additional insights
    generate_insights(student_data, title_suffix)

    # Preprocess the data for machine learning
    X = student_data.drop('G3', axis=1).select_dtypes(include=[np.number])
    y = student_data['G3']

    # Split the data into training and test sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

    # Train a linear regression model
    model = LinearRegression()
    model.fit(X_train, y_train)

    # Make predictions
    predictions = model.predict(X_test)

    # Evaluate the model
    mse = mean_squared_error(y_test, predictions)
    r2 = r2_score(y_test, predictions)
```

```
    print(f"Mean Squared Error for {title_suffix}: {mse}")
    print(f"R^2 Score for {title_suffix}: {r2}")

    # Plot the true vs predicted values
    plt.figure(figsize=(10, 6))
    plt.scatter(y_test, predictions)
    plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red',
linestyle='--')
    plt.title(f'True vs Predicted Final Grades (G3) ({title_suffix})')
    plt.xlabel('True Grades (G3)')
    plt.ylabel('Predicted Grades (G3)')
    plt.show()

# Function to generate insights from the correlation matrix
def generate_insights_from_correlation(correlation_matrix, title_suffix):
    strong_correlations = correlation_matrix[(correlation_matrix > 0.5) |
(correlation_matrix < -0.5)]
    strong_correlations = strong_correlations.dropna(how='all',
axis=0).dropna(how='all', axis=1)
    print(f"Strong Correlations in {title_suffix} Data:\n", strong_correlations)

    # Generate some heuristic-based insights
    warnings = []
    suggestions = []
    positives = []
    negatives = []

    for col in strong_correlations.columns:
        for idx in strong_correlations.index:
            if col != idx and not np.isnan(strong_correlations.loc[idx, col]):
                corr_value = strong_correlations.loc[idx, col]
                if corr_value > 0.7:
                    positives.append(f"High positive correlation ({corr_value:.2f})
between {col} and {idx}.")
                elif corr_value < -0.7:
                    negatives.append(f"High negative correlation ({corr_value:.2f})
between {col} and {idx}.")
                elif corr_value > 0.5:
                    suggestions.append(f"Moderate positive correlation
({corr_value:.2f}) between {col} and {idx}.")
                elif corr_value < -0.5:
                    warnings.append(f"Moderate negative correlation
({corr_value:.2f}) between {col} and {idx}.")

    # Print insights
    print(f"\nWarnings for {title_suffix} Data:")
    for warning in warnings:
        print(warning)
```

```
    print(f"\nSuggestions for {title_suffix} Data:")
    for suggestion in suggestions:
        print(suggestion)

    print(f"\nPositives for {title_suffix} Data:")
    for positive in positives:
        print(positive)

    print(f"\nNegatives for {title_suffix} Data:")
    for negative in negatives:
        print(negative)

# Function to generate additional insights
def generate_insights(student_data, title_suffix):
    # Generate some heuristic-based insights
    warnings = []
    suggestions = []
    positives = []
    negatives = []

    # Impact of study time on final grades
    avg_grades_by_studytime = student_data.groupby('studytime')['G3'].mean()
    if avg_grades_by_studytime.idxmax() > 2:
        positives.append("Students who study more than 2 hours tend to have higher
final grades.")
    else:
        suggestions.append("Encourage students to increase study time to improve
final grades.")

    # Impact of weekday alcohol consumption on final grades
    avg_grades_by_dalc = student_data.groupby('Dalc')['G3'].mean()
    if avg_grades_by_dalc.idxmax() > 2:
        warnings.append("High weekday alcohol consumption is associated with lower
final grades.")
    else:
        positives.append("Low weekday alcohol consumption is associated with higher
final grades.")

    # Impact of weekend alcohol consumption on final grades
    avg_grades_by_walc = student_data.groupby('Walc')['G3'].mean()
    if avg_grades_by_walc.idxmax() > 2:
        warnings.append("High weekend alcohol consumption is associated with lower
final grades.")
    else:
        positives.append("Low weekend alcohol consumption is associated with higher
final grades.")
```

```
    # Print insights
    print(f"\nAdditional Warnings for {title_suffix} Data:")
    for warning in warnings:
        print(warning)

    print(f"\nAdditional Suggestions for {title_suffix} Data:")
    for suggestion in suggestions:
        print(suggestion)

    print(f"\nAdditional Positives for {title_suffix} Data:")
    for positive in positives:
        print(positive)

    print(f"\nAdditional Negatives for {title_suffix} Data:")
    for negative in negatives:
        print(negative)

# Analyze the student performance data for mathematics
analyze_student_data(student_mat_data, "Mathematics")

# Analyze the student performance data for Portuguese
analyze_student_data(student_por_data, "Portuguese")
```

Dataset output of the following code

Summary Statistics for Mathematics:

	age	Medu	Fedu	traveltime	studytime	failures	famrel	...	Dalc	Walc	health	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	...	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304	...	1.481013	2.291139	3.554430	5.708861	10.908861	10.713924	10.415190
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	0.896659	...	0.890741	1.287897	1.390303	8.003096	3.319195	3.761505	4.581443
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	...	1.000000	1.000000	1.000000	0.000000	3.000000	0.000000	0.000000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	4.000000	...	1.000000	1.000000	3.000000	0.000000	8.000000	9.000000	8.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000	...	1.000000	2.000000	4.000000	4.000000	11.000000	11.000000	11.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	...	2.000000	3.000000	5.000000	8.000000	13.000000	13.000000	14.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	...	5.000000	5.000000	5.000000	75.000000	19.000000	19.000000	20.000000

[8 rows x 16 columns]

Strong Correlations in Mathematics Data:

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Medu	NaN	1.000000	0.623455	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fedu	NaN	0.623455	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
traveltime	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
studytime	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
failures	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
famrel	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freetime	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
goout	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Dalc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.647544	NaN	NaN	NaN	NaN	NaN
Walc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.647544	1.000000	NaN	NaN	NaN	NaN	NaN
health	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN
absences	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN
G1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.852118	0.801468
G2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.852118	1.000000	0.904868
G3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.801468	0.904868	1.000000

Suggestions for Mathematics Data:

Moderate positive correlation (0.62) between Medu and Fedu.

Moderate positive correlation (0.62) between Fedu and Medu.

Moderate positive correlation (0.65) between Dalc and Walc.

Moderate positive correlation (0.65) between Walc and Dalc.

Positives for Mathematics Data:

High positive correlation (0.85) between G1 and G2.

High positive correlation (0.80) between G1 and G3.

High positive correlation (0.85) between G2 and G1.

High positive correlation (0.90) between G2 and G3.

High positive correlation (0.80) between G3 and G1.

High positive correlation (0.90) between G3 and G2.

Additional Positives for Mathematics Data:

Students who study more than 2 hours tend to have higher final grades.

Low weekday alcohol consumption is associated with higher final grades.

Low weekend alcohol consumption is associated with higher final grades.

Additional Negatives for Mathematics Data:

Mean Squared Error for Mathematics: 4.503769153719581

R^2 Score for Mathematics: 0.7803580213768332

Summary Statistics for Portuguese:

	age	Medu	Fedu	traveltime	studytime	failures	famrel	...	Dalc	Walc	health	absences	G1	G2	G3
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	...	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	16.744222	2.514638	2.306626	1.568567	1.930663	0.221880	3.930663	...	1.502311	2.280431	3.536210	3.659476	11.399076	11.570108	11.906009
std	1.218138	1.134552	1.099931	0.748660	0.829510	0.593235	0.955717	...	0.924834	1.284380	1.446259	4.640759	2.745265	2.913639	3.230656
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	...	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000	4.000000	...	1.000000	1.000000	2.000000	0.000000	10.000000	10.000000	10.000000
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000	4.000000	...	1.000000	2.000000	4.000000	2.000000	11.000000	11.000000	12.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	...	2.000000	3.000000	5.000000	6.000000	13.000000	13.000000	14.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	...	5.000000	5.000000	5.000000	32.000000	19.000000	19.000000	19.000000

[8 rows x 16 columns]

Strong Correlations in Portuguese Data:

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Medu	NaN	1.000000	0.647477	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fedu	NaN	0.647477	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
traveltime	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
studytime	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
failures	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
famrel	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freetime	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
goout	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Dalc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.616561	NaN	NaN	NaN	NaN	NaN
Walc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.616561	1.000000	NaN	NaN	NaN	NaN	NaN
health	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN
absences	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN
G1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.864982	0.826387
G2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.864982	1.000000	0.918548
G3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.826387	0.918548	1.000000

Suggestions for Portuguese Data:

Moderate positive correlation (0.65) between Medu and Fedu.

Moderate positive correlation (0.65) between Fedu and Medu.

Moderate positive correlation (0.62) between Dalc and Walc.

Moderate positive correlation (0.62) between Walc and Dalc.

Positives for Portuguese Data:

High positive correlation (0.86) between G1 and G2.

High positive correlation (0.83) between G1 and G3.

High positive correlation (0.86) between G2 and G1.

absences	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN
G1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.864982	0.826387
G2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.864982	1.000000	0.918548
G3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.826387	0.918548	1.000000

Suggestions for Portuguese Data:

Moderate positive correlation (0.65) between Medu and Fedu.

Moderate positive correlation (0.65) between Fedu and Medu.

Moderate positive correlation (0.62) between Dalc and Walc.

Moderate positive correlation (0.62) between Walc and Dalc.

Positives for Portuguese Data:

High positive correlation (0.86) between G1 and G2.

High positive correlation (0.83) between G1 and G3.

High positive correlation (0.86) between G2 and G1.

Warnings for Portuguese Data:

Suggestions for Portuguese Data:

Moderate positive correlation (0.65) between Medu and Fedu.

Moderate positive correlation (0.65) between Fedu and Medu.

Moderate positive correlation (0.62) between Dalc and Walc.

Moderate positive correlation (0.62) between Walc and Dalc.

Positives for Portuguese Data:

High positive correlation (0.86) between G1 and G2.

High positive correlation (0.83) between G1 and G3.

High positive correlation (0.86) between G2 and G1.

Suggestions for Portuguese Data:

Moderate positive correlation (0.65) between Medu and Fedu.

Moderate positive correlation (0.65) between Fedu and Medu.

Moderate positive correlation (0.62) between Dalc and Walc.

Moderate positive correlation (0.62) between Walc and Dalc.

Positives for Portuguese Data:

High positive correlation (0.86) between G1 and G2.

High positive correlation (0.83) between G1 and G3.

High positive correlation (0.86) between G2 and G1.

Moderate positive correlation (0.62) between Dalc and Walc.

Moderate positive correlation (0.62) between Walc and Dalc.

Positives for Portuguese Data:

High positive correlation (0.86) between G1 and G2.

High positive correlation (0.83) between G1 and G3.

High positive correlation (0.86) between G2 and G1.

High positive correlation (0.92) between G2 and G3.

High positive correlation (0.83) between G3 and G1.

Positives for Portuguese Data:

High positive correlation (0.86) between G1 and G2.

High positive correlation (0.83) between G1 and G3.

High positive correlation (0.86) between G2 and G1.

High positive correlation (0.92) between G2 and G3.

High positive correlation (0.83) between G3 and G1.

High positive correlation (0.92) between G3 and G2.

Additional Positives for Portuguese Data:

Students who study more than 2 hours tend to have higher final grades.

Low weekday alcohol consumption is associated with higher final grades.

Low weekend alcohol consumption is associated with higher final grades.

Additional Negatives for Portuguese Data:

Mean Squared Error for Portuguese: 1.3605829329063632

Additional Positives for Portuguese Data:

Students who study more than 2 hours tend to have higher final grades.

Low weekday alcohol consumption is associated with higher final grades.

Low weekend alcohol consumption is associated with higher final grades.

Additional Negatives for Portuguese Data:

Mean Squared Error for Portuguese: 1.3605829329063632

Low weekday alcohol consumption is associated with higher final grades.

Low weekend alcohol consumption is associated with higher final grades.

Additional Negatives for Portuguese Data:

Mean Squared Error for Portuguese: 1.3605829329063632

Mean Squared Error for Portuguese: 1.3605829329063632

R^2 Score for Portuguese: 0.8604775881282157

STUDIES AND GRAPHS

LIST OF VARIOUS GRAPHS GENERATED

MATHEMETICS

1. CORELATION MATRIX OF STUDENT PERFORMANCE DATA (MATHEMETICS)

2. DISTRIBUTION OF FINAL GRADES G3 (MATHEMETICS)

3. IMPACT OF STUDY TIME IN FINAL GRADES G3 (MATHEMETICS)

4. IMPACT OF WEEKDAY ALCOHOL CONSUMPTIONS IN FINAL GRADES G3 (MATHEMETICS)

5. IMPACT OF WEEKEND ALCOHOL CONSUMPTIONS IN FINAL GRADES G3 (MATHEMETICS)

PORTUGUESE

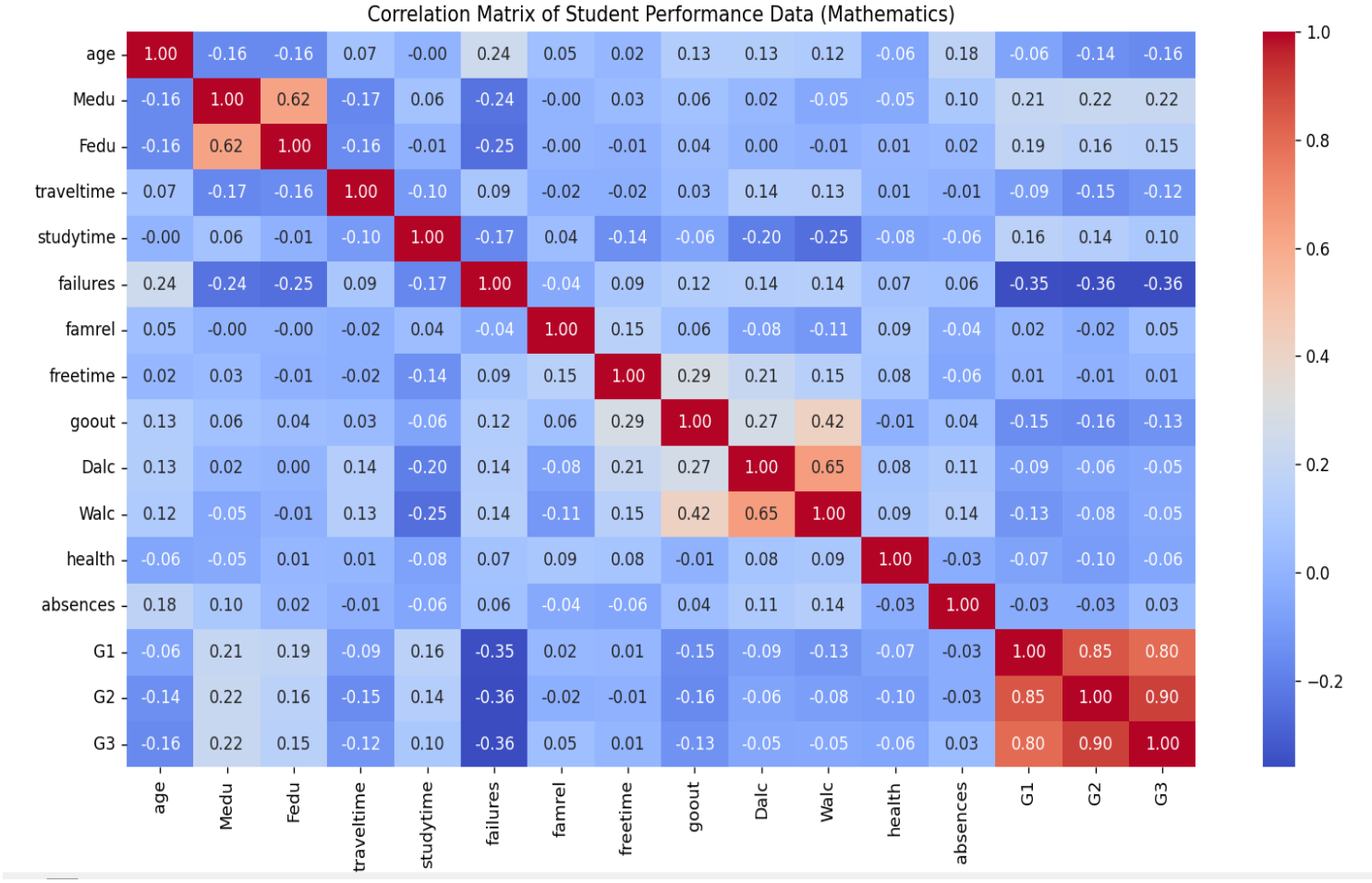
1. CORELATION MATRIX OF STUDENT PERFORMANCE DATA (PORTUGUESE)

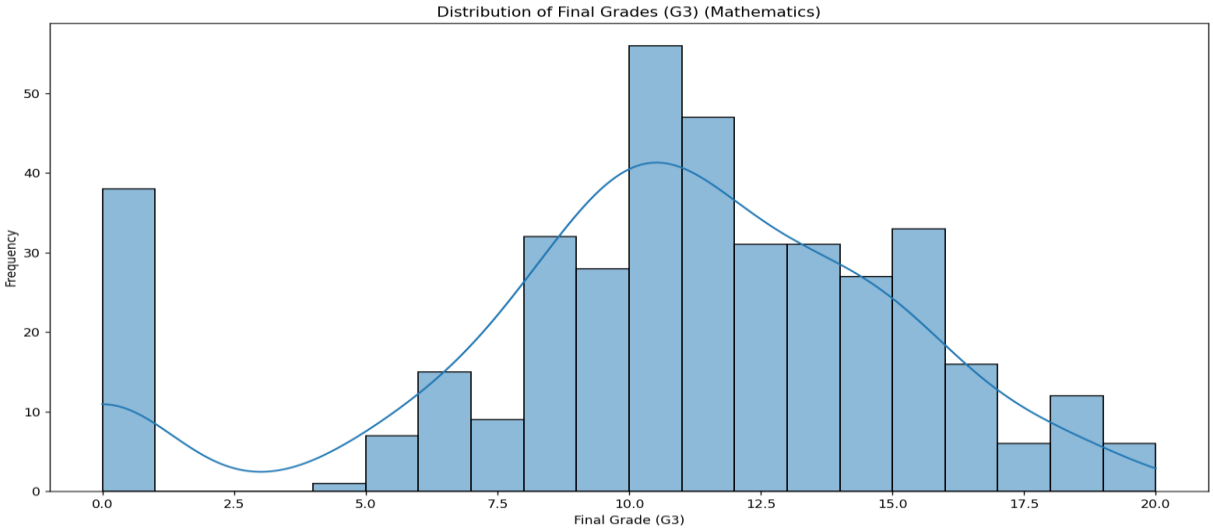
2. DISTRIBUTION OF FINAL GRADES G3 (PORTUGUESE)

3. IMPACT OF STUDY TIME IN FINAL GRADES G3 (PORTUGUESE)

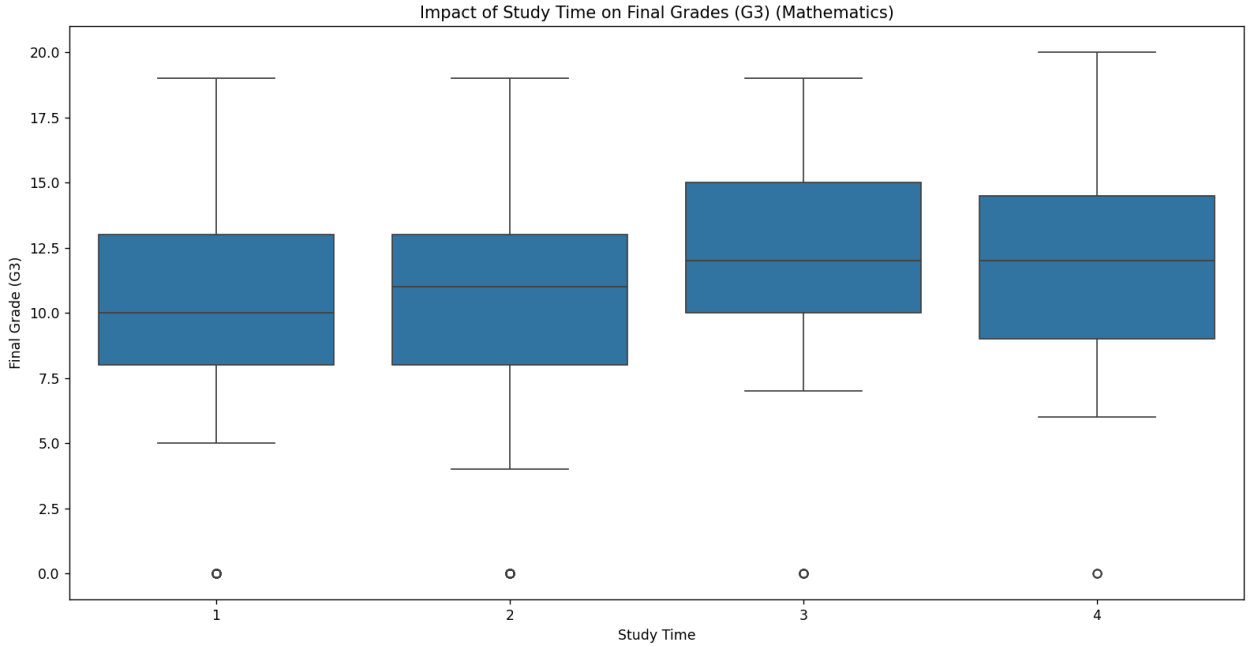
4. IMPACT OF WEEKDAY ALCOHOL CONSUMPTIONS IN FINAL GRADES G3 (PORTUGUESE)

5. IMPACT OF WEEKEND ALCOHOL CONSUMPTIONS IN FINAL GRADES G3 (PORTUGUESE)

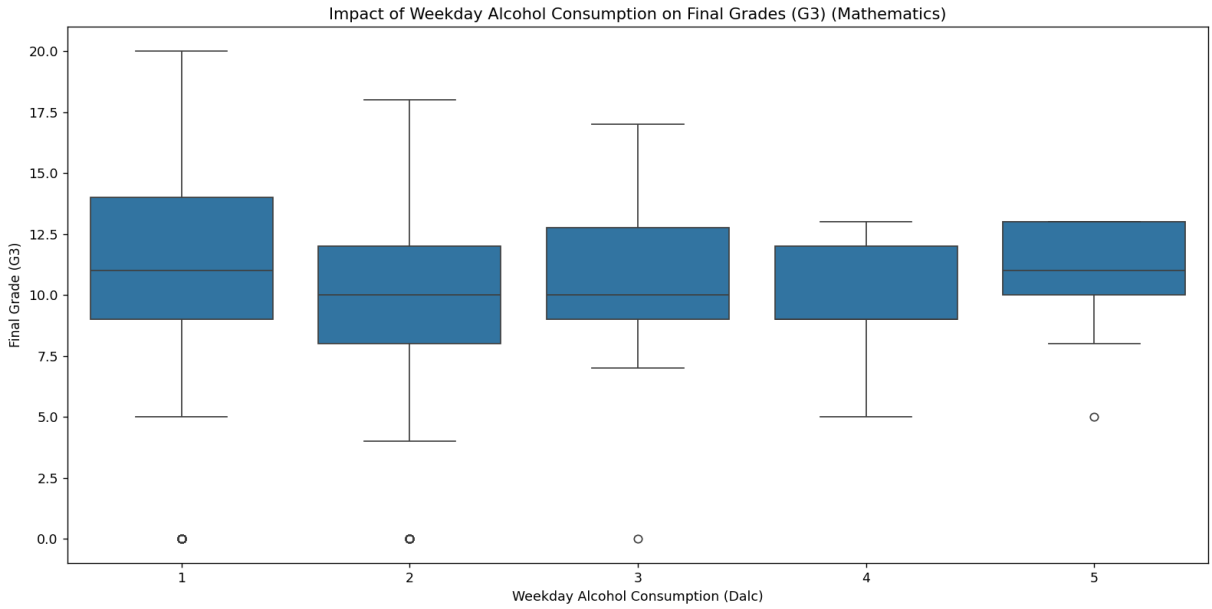

Correlation Matrix of Student Performance Data (Mathematics)
age 1.00 -0.16 -0.16 0.07 -0.00 0.24 0.05 0.02 0.13 0.13 0.12 -0.06 0.18 -0.06 -0.14 -0.16
Medu -0.16 1.00 0.62 -0.17 0.06 -0.24 -0.00 0.03 0.06 0.02 -0.05 -0.05 0.10 0.21 0.22 0.22
Fedu -0.16 0.62 1.00 -0.16 -0.01 -0.25 -0.00 -0.01 0.04 0.00 -0.01 0.01 0.02 0.19 0.16 0.15
traveltime 0.07 -0.17 -0.16 1.00 -0.10 0.09 -0.02 -0.02 0.03 0.14 0.13 0.01 -0.01 -0.09 -0.15 -0.12
studytime -0.00 0.06 -0.01 -0.10 1.00 -0.17 0.04 -0.14 -0.06 -0.20 -0.25 -0.08 -0.06 0.16 0.14 0.10
failures 0.24 -0.24 -0.25 0.09 -0.17 1.00 -0.04 0.09 0.12 0.14 0.14 0.07 0.06 -0.35 -0.36 -0.36
famrel 0.05 -0.00 -0.00 -0.02 0.04 -0.04 1.00 0.15 0.06 -0.08 -0.11 0.09 -0.04 0.02 -0.02 0.05
freetime 0.02 0.03 -0.01 -0.02 -0.14 0.09 0.15 1.00 0.29 0.21 0.15 0.08 -0.06 0.01 -0.01 0.01
goout 0.13 0.06 0.04 0.03 -0.06 0.12 0.06 0.29 1.00 0.27 0.42 -0.01 0.04 -0.15 -0.16 -0.13
Dalc 0.13 0.02 0.00 0.14 -0.20 0.14 -0.08 0.21 0.27 1.00 0.65 0.08 0.11 -0.09 -0.06 -0.05
Walc 0.12 -0.05 -0.01 0.13 -0.25 0.14 -0.11 0.15 0.42 0.65 1.00 0.09 0.14 -0.13 -0.08 -0.05
health -0.06 -0.05 0.01 0.01 -0.08 0.07 0.09 0.08 -0.01 0.08 0.09 1.00 -0.03 -0.07 -0.10 -0.06
absences 0.18 0.10 0.02 -0.01 -0.06 0.06 -0.04 -0.06 0.04 0.11 0.14 -0.03 1.00 -0.03 -0.03 0.03
G1 -0.06 0.21 0.19 -0.09 0.16 -0.35 0.02 0.01 -0.15 -0.09 -0.13 -0.07 -0.03 1.00 0.85 0.80
G2 -0.14 0.22 0.16 -0.15 0.14 -0.36 -0.02 -0.01 -0.16 -0.06 -0.08 -0.10 -0.03 0.85 1.00 0.90
G3 -0.16 0.22 0.15 -0.12 0.10 -0.36 0.05 0.01 -0.13 -0.05 -0.05 -0.06 0.03 0.80 0.90 1.00
age
Medu
Fedu
traveltime
studytime
failures
famrel
freetime
goout
Dalc
Walc
health
absences
G1
G2
G3
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Distribution of Final Grades (G3) (Mathematics)
Frequency
Final Grade (G3)
0
10
20
30
40
50
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

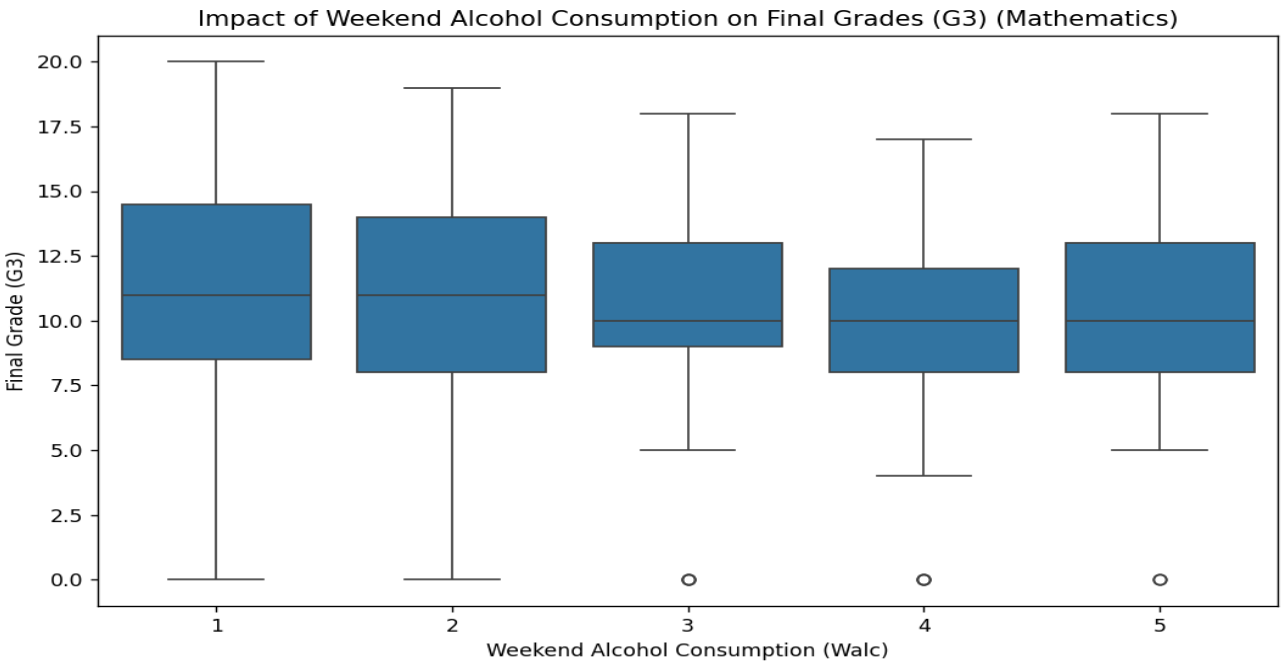
2.


Impact of Study Time on Final Grades (G3) (Mathematics)
Final Grade (G3)
Study Time
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
1
2
3
4

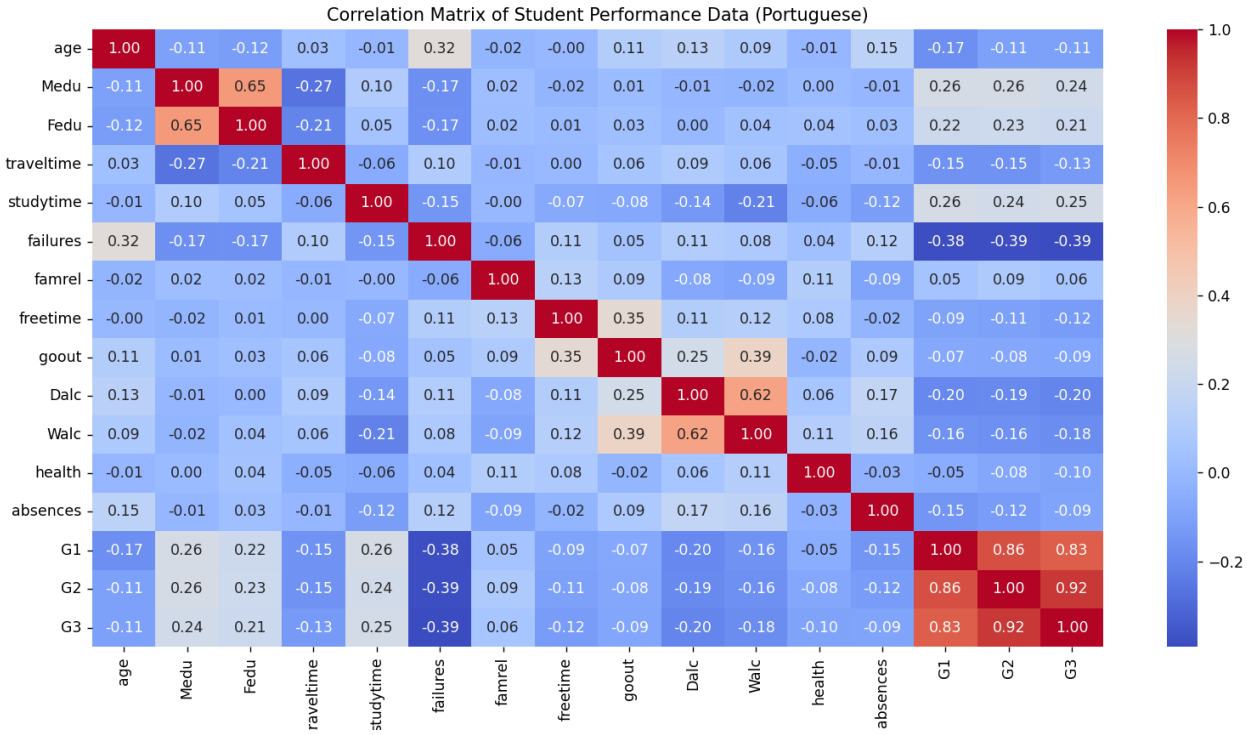
3.


Impact of Weekday Alcohol Consumption on Final Grades (G3) (Mathematics)
Final Grade (G3)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Weekday Alcohol Consumption (Dalc)

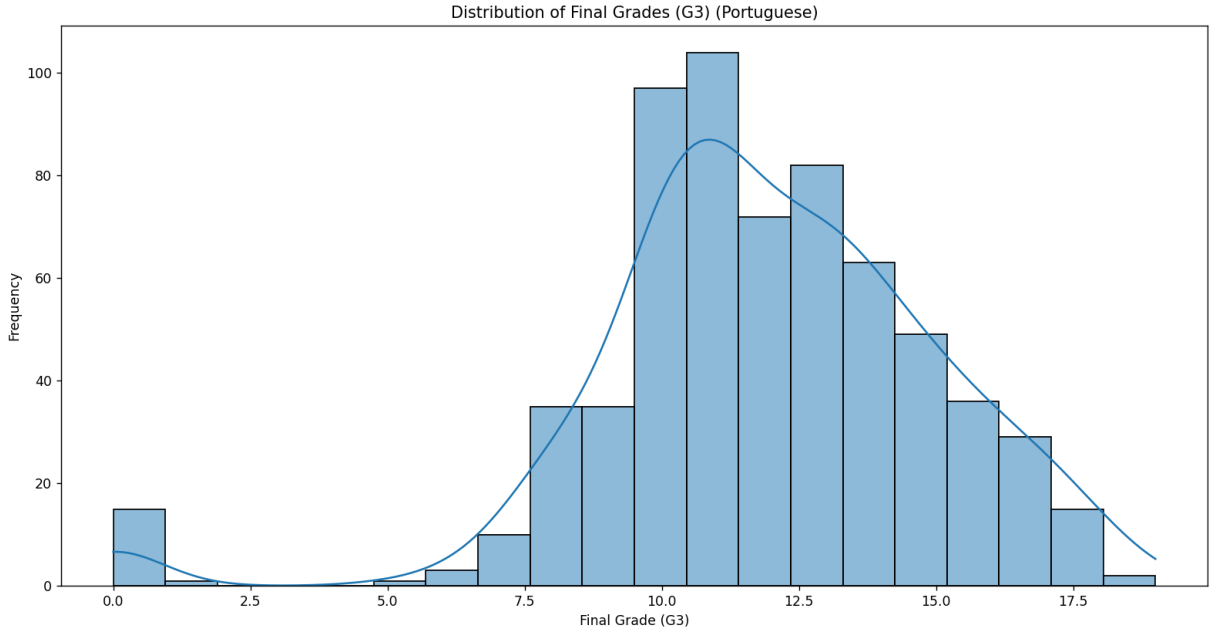
4.


Impact of Weekend Alcohol Consumption on Final Grades (G3) (Mathematics)
Final Grade (G3)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Weekend Alcohol Consumption (Walc)

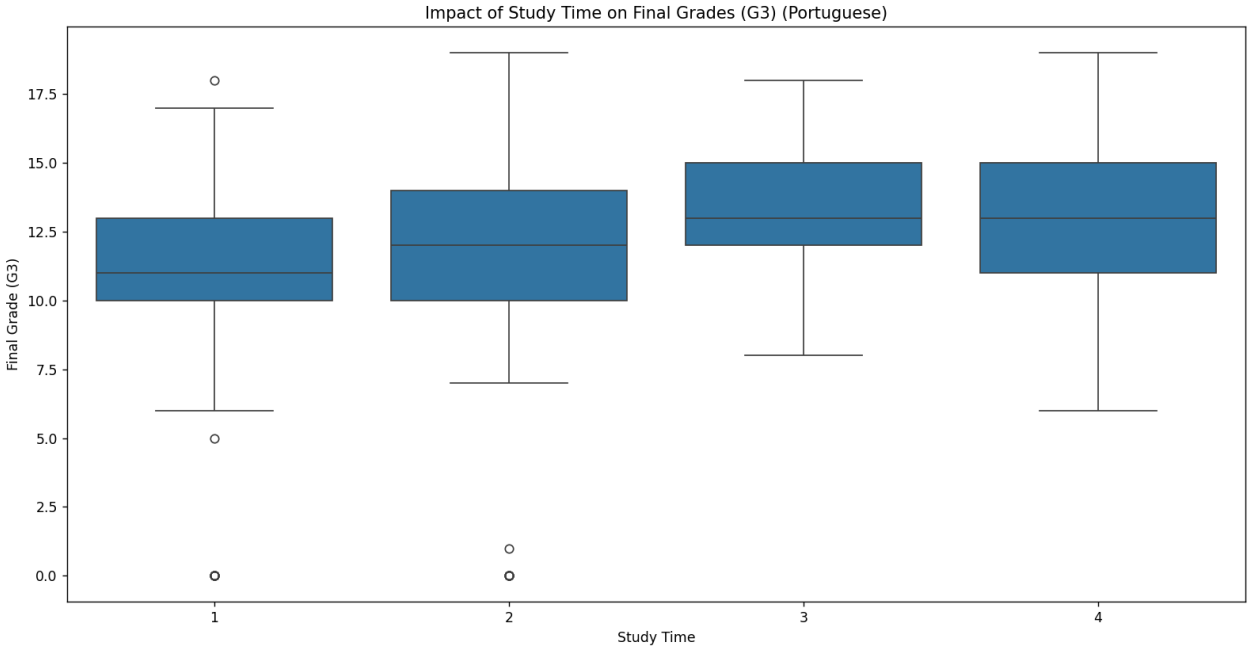
5.


Correlation Matrix of Student Performance Data (Portuguese)
age 1.00 -0.11 -0.12 0.03 -0.01 0.32 -0.02 -0.00 0.11 0.13 0.09 -0.01 0.15 -0.17 -0.11 -0.11
Medu -0.11 1.00 0.65 -0.27 0.10 -0.17 0.02 -0.02 0.01 -0.01 -0.02 0.00 -0.01 0.26 0.26 0.24
Fedu -0.12 0.65 1.00 -0.21 0.05 -0.17 0.02 0.01 0.03 0.00 0.04 0.04 0.03 0.22 0.23 0.21
traveltime 0.03 -0.27 -0.21 1.00 -0.06 0.10 -0.01 0.00 0.06 0.09 0.06 -0.05 -0.01 -0.15 -0.15 -0.13
studytime -0.01 0.10 0.05 -0.06 1.00 -0.15 -0.00 -0.07 -0.08 -0.14 -0.21 -0.06 -0.12 0.26 0.24 0.25
failures 0.32 -0.17 -0.17 0.10 -0.15 1.00 -0.06 0.11 0.05 0.11 0.08 0.04 0.12 -0.38 -0.39 -0.39
famrel -0.02 0.02 0.02 -0.01 -0.00 -0.06 1.00 0.13 0.09 -0.08 -0.09 0.11 -0.09 0.05 0.09 0.06
freetime -0.00 -0.02 0.01 0.00 -0.07 0.11 0.13 1.00 0.35 0.11 0.12 0.08 -0.02 -0.09 -0.11 -0.12
goout 0.11 0.01 0.03 0.06 -0.08 0.05 0.09 0.35 1.00 0.25 0.39 -0.02 0.09 -0.07 -0.08 -0.09
Dalc 0.13 -0.01 0.00 0.09 -0.14 0.11 -0.08 0.11 0.25 1.00 0.62 0.06 0.17 -0.20 -0.19 -0.20
Walc 0.09 -0.02 0.04 0.06 -0.21 0.08 -0.09 0.12 0.39 0.62 1.00 0.11 0.16 -0.16 -0.16 -0.18
health -0.01 0.00 0.04 -0.05 -0.06 0.04 0.11 0.08 -0.02 0.06 0.11 1.00 -0.03 -0.05 -0.08 -0.10
absences 0.15 -0.01 0.03 -0.01 -0.12 0.12 -0.09 -0.02 0.09 0.17 0.16 -0.03 1.00 -0.15 -0.12 -0.09
G1 -0.17 0.26 0.22 -0.15 0.26 -0.38 0.05 -0.09 -0.07 -0.20 -0.16 -0.05 -0.15 1.00 0.86 0.83
G2 -0.11 0.26 0.23 -0.15 0.24 -0.39 0.09 -0.11 -0.08 -0.19 -0.16 -0.08 -0.12 0.86 1.00 0.92
G3 -0.11 0.24 0.21 -0.13 0.25 -0.39 0.06 -0.12 -0.09 -0.20 -0.18 -0.10 -0.09 0.83 0.92 1.00
age
Medu
Fedu
traveltime
studytime
failures
famrel
freetime
goout
Dalc
Walc
health
absences
G1
G2
G3
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

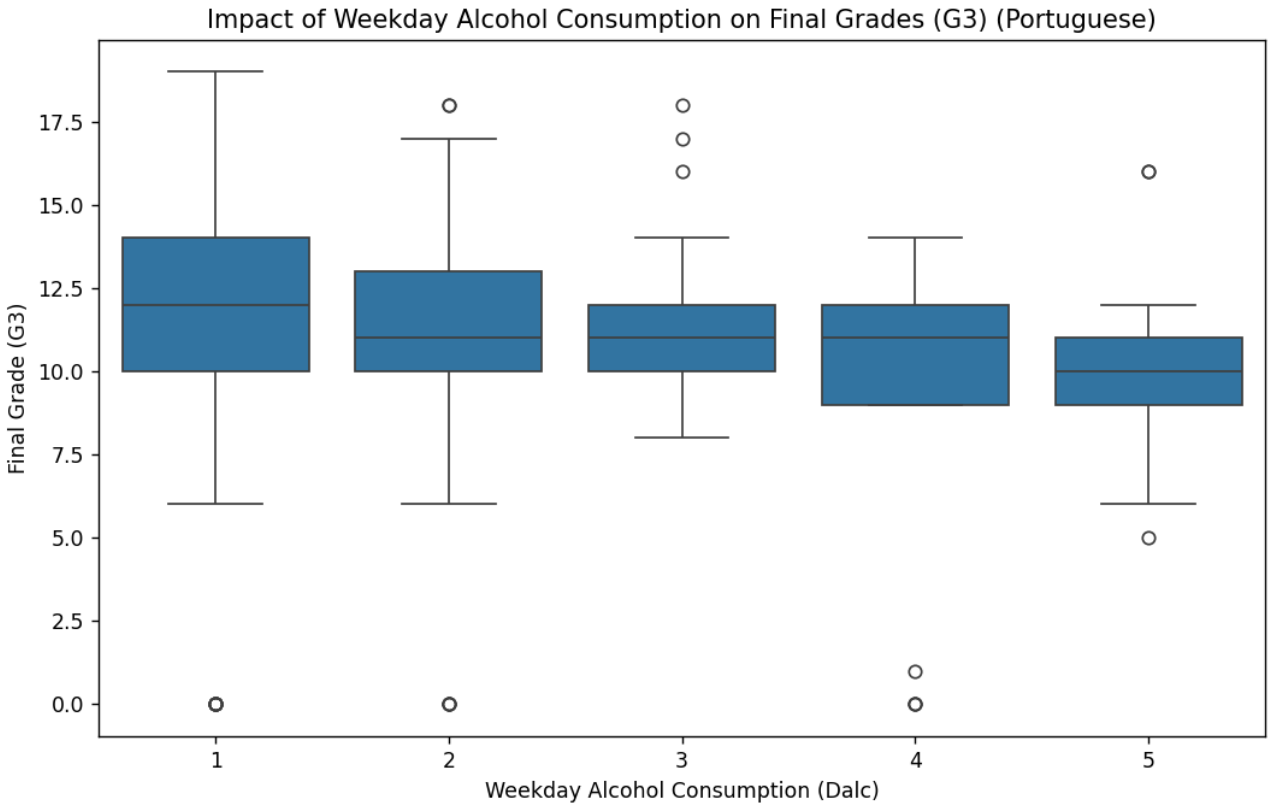
1.


Distribution of Final Grades (G3) (Portuguese)
Frequency
0
20
40
60
80
100
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Final Grade (G3)

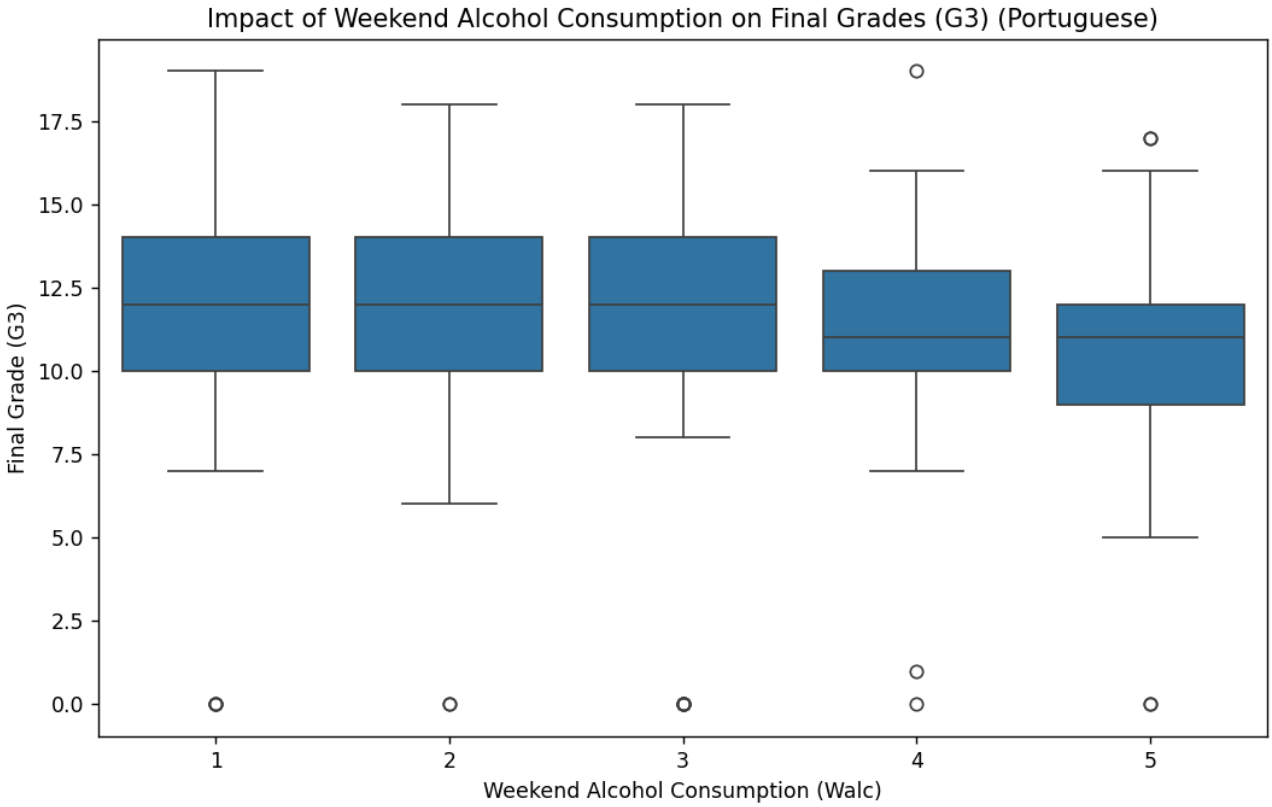
2.


Impact of Study Time on Final Grades (G3) (Portuguese)
Final Grade (G3)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
Study Time

3.

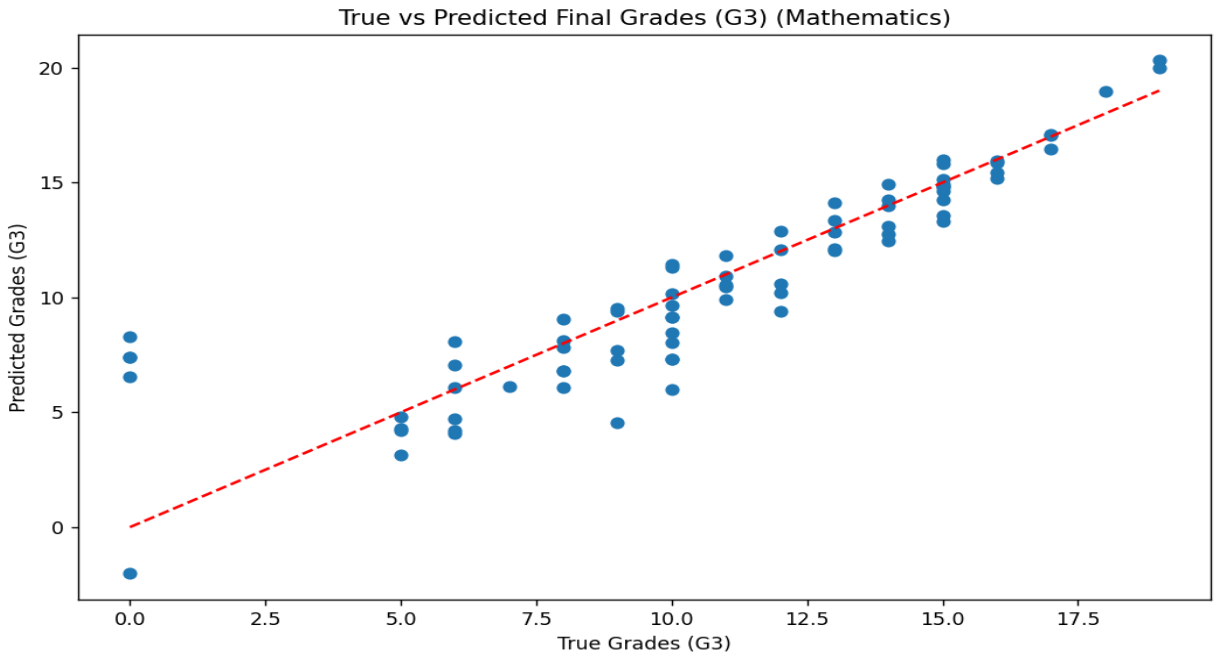

Impact of Weekday Alcohol Consumption on Final Grades (G3) (Portuguese)
Final Grade (G3)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Weekday Alcohol Consumption (Dalc)

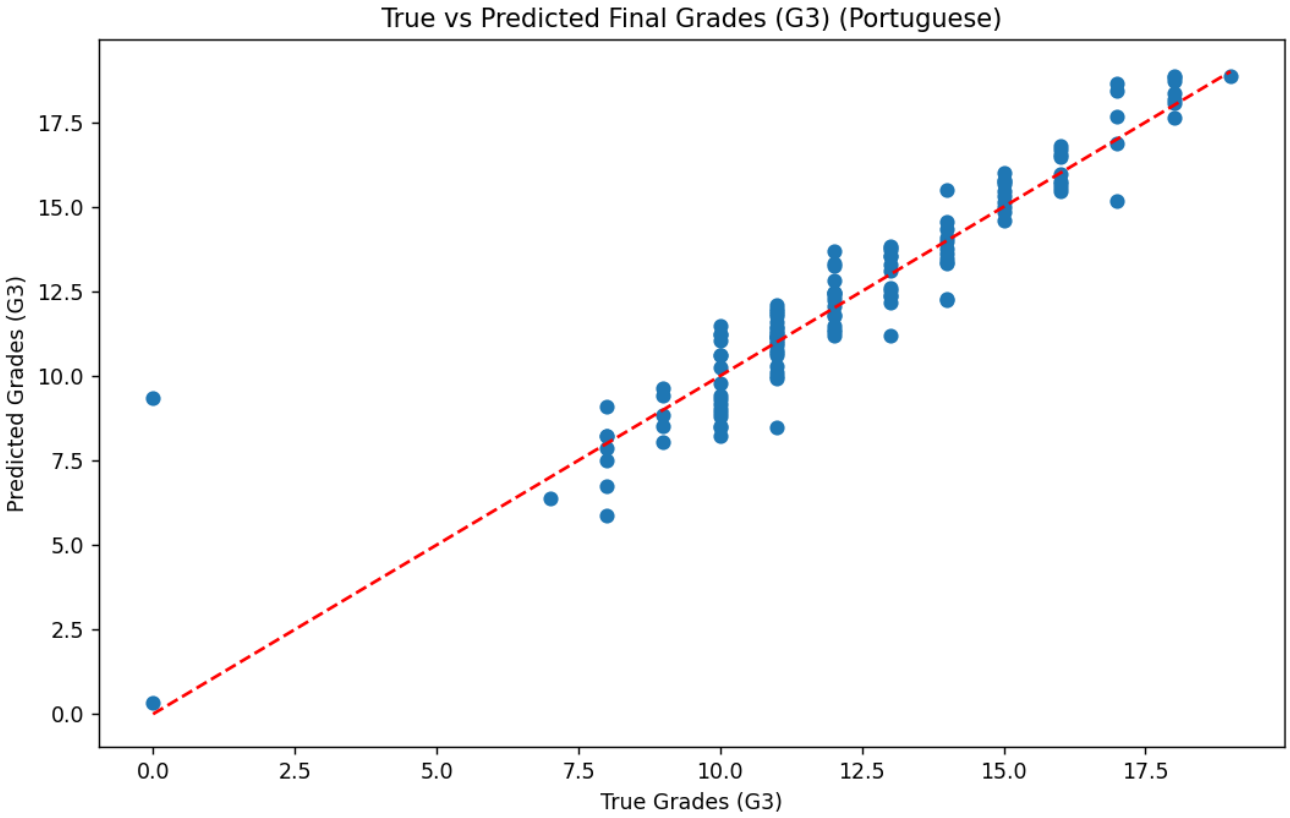
4.



5.

TRUE VERSUS PREDICT DATASET GRAPH




True vs Predicted Final Grades (G3) (Portuguese)
Predicted Grades (G3)
True Grades (G3)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

**********THANKYOU**********