MTMC-512 Programming Lab IV (Machine Learning)

Lab Assignment 3: Exploratory Data Analysis - Student Performance Dataset

1 Task 1: Load and Explore the Dataset

1.1 1.1 Load the Dataset

```
import pandas as pd
df = pd.read_csv(’student-mat.csv’, sep=’;’)
df.head()
```

```
  school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason ... G1 G2 G3
0  GP     F   18   U       GT3     A        4    4    at   teacher course ... 5 6 6
1  GP     F   17   U       GT3     T        1    1    at   other   course ... 5 5 6
2  GP     F   15   U       LE3     T        1    1    at   other   other  ... 7 8 10
3  GP     F   15   U       GT3     T        4    2    health services course ... 15 14 15
4  GP     F   16   U       GT3     T        3    3    other other    home   ... 6 10 10
```

1.2 1.2 Dataset Characteristics

Number of Records and Features:

```
df.shape
```

```
(395, 33)
```

Data Types of Columns:

```
df.dtypes
```

```
school     object
sex        object
age         int64
address    object
famsize    object
... (remaining columns)
G3          int64
dtype: object
```

Summary Statistics:

```
df.describe(include='all')
```

```
        school  sex        age   address  ...     G1     G2     G3
count     395   395        395      395   ...    395    395    395
unique      2     2        NaN        2   ...    NaN    NaN    NaN
top        GP     M        NaN        U   ...    NaN    NaN    NaN
freq      307   208        NaN      307   ...    NaN    NaN    NaN
mean      NaN   NaN    16.696    NaN     ...  10.91  10.71  10.42
```

1.3 1.3 Check for Missing Values, Duplicates, and Outliers

Missing Values:

```
df.isnull().sum()
```

```
All columns: 0 missing values
```

Duplicates:

```
df.duplicated().sum()
```

```
0
```

Outliers (Boxplot of Grades):

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12,6))
sns.boxplot(data=df[['G1', 'G2', 'G3']])
plt.title("Boxplot of Grades")
plt.savefig("boxplot_grades.png")
plt.show()
```

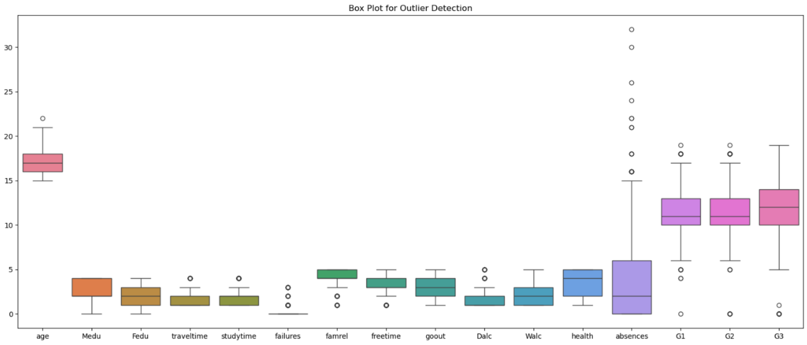

Figure 1: Boxplot for Student Grades

2 Task 2: Statistical Analysis

2.1 2.1 Basic Statistics

```
df[['G1','G2','G3']].agg(['mean', 'median', 'var', 'skew', 'kurt'])
```

```
             G1        G2        G3
mean     10.91     10.71     10.42
median   11.00     11.00     11.00
var      7.30      7.41      11.00
skew     -0.19     -0.10     -0.25
kurt     -0.67     -0.64     -0.39
```

2.2 2.2 Correlation Analysis

```python
correlation_matrix = df.corr(numeric_only=True)
correlation_matrix['G3'].sort_values(ascending=False)
```

```
G3           1.000000
G2           0.904868
G1           0.852119
failures    -0.360415
absences    -0.053929
studytime    0.097820
Name: G3, dtype: float64
```

2.3 2.3 Correlation Matrix Heatmap

```python
plt.figure(figsize=(14,10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.savefig("heatmap_correlation.png")
plt.show()
```

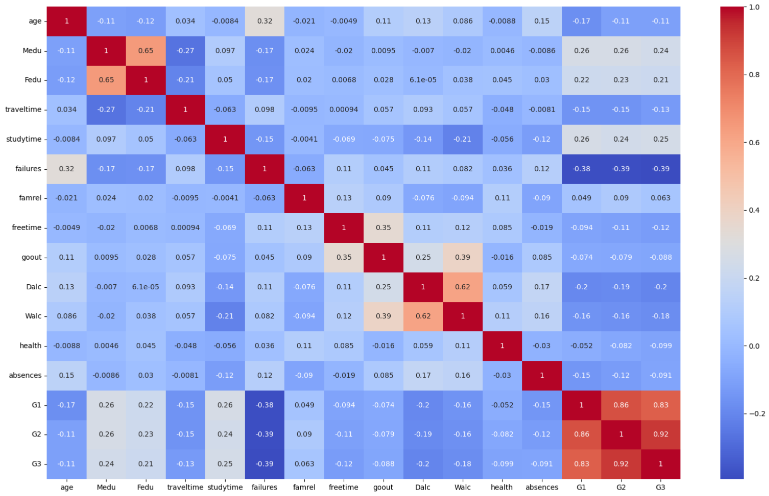

Figure 2: Correlation Heatmap

3 Task 3: Data Visualization

3.1 3.1 Univariate Analysis

```python
# Create a figure with 1 row and 3 columns
plt.figure(figsize=(18, 5))

# Histogram for Exam Scores
plt.subplot(1, 3, 1)
sns.histplot(data=df, x='G1', kde=True, bins=20, color='skyblue')
plt.title('Exam Scores Distribution')
```

```

# Histogram for Study Time
plt.subplot(1, 3, 2)
sns.histplot(data=df, x=’studytime’, kde=True, bins=20, color=’salmon’)
plt.title(’Study Time Distribution’)

# Histogram for Attendance
plt.subplot(1, 3, 3)
sns.histplot(data=df, x=’absences’, kde=True, bins=20, color=’lightgreen’)
plt.title(’Attendance Distribution’)

plt.tight_layout()
plt.savefig(’Histograms’)
plt.show()
```

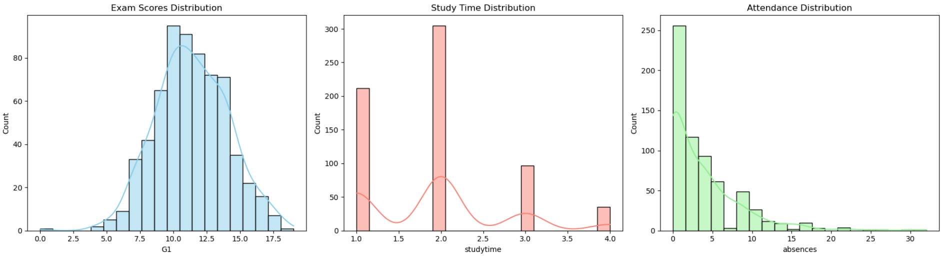

Figure 3: Histograms of Exam Scores, Study Time, and Attendance

3.2 3.2 Bivariate Analysis

```
sns.scatterplot(x=’studytime’, y=’G3’, data=df)
plt.title("Study Time vs Final Grade (G3)")
plt.savefig("scatter_studytime_g3.png")
plt.show()
```

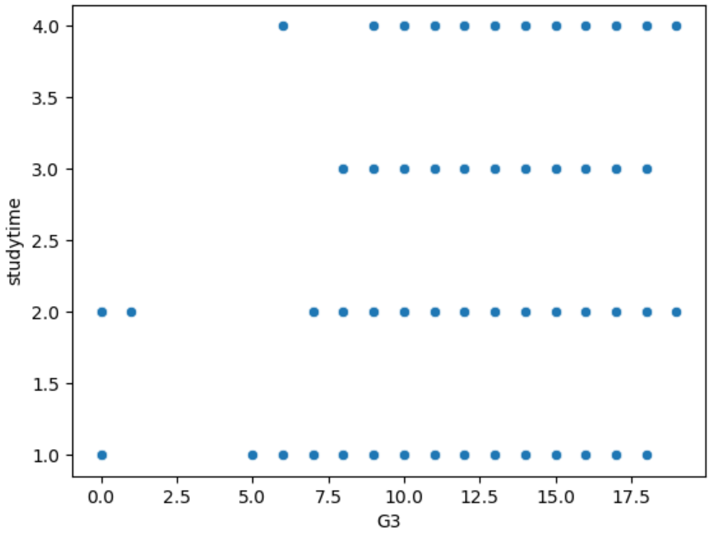

Figure 4: Scatter Plot: Study Time vs G3

Pairplot:

```
sns.pairplot(df[['G1','G2','G3','studytime','absences']])
plt.savefig("pairplot_students.png")
plt.show()
```

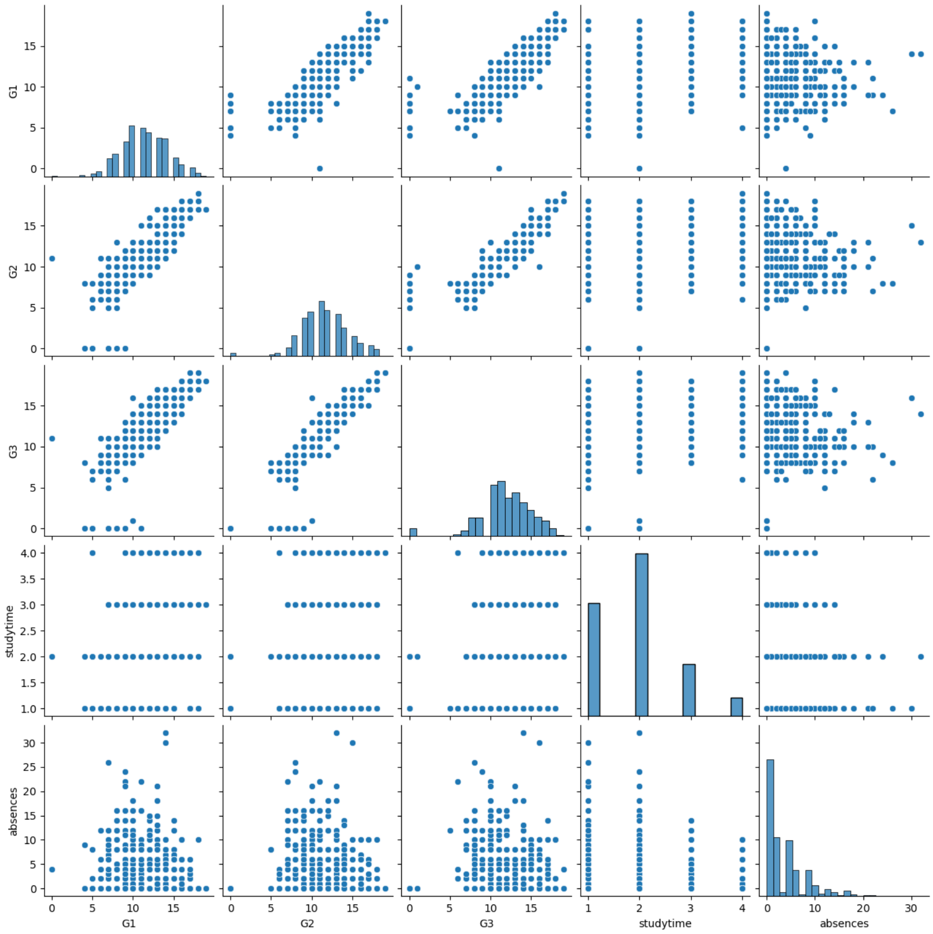

Figure 5: Pairplot of Selected Features

3.3 3.3 Categorical Data Visualization

Parental Education vs Performance:

```
sns.barplot(x='Medu', y='G3', data=df)
plt.title("Mother's Education vs Final Grade")
plt.savefig("bar_medu_g3.png")
plt.show()
```

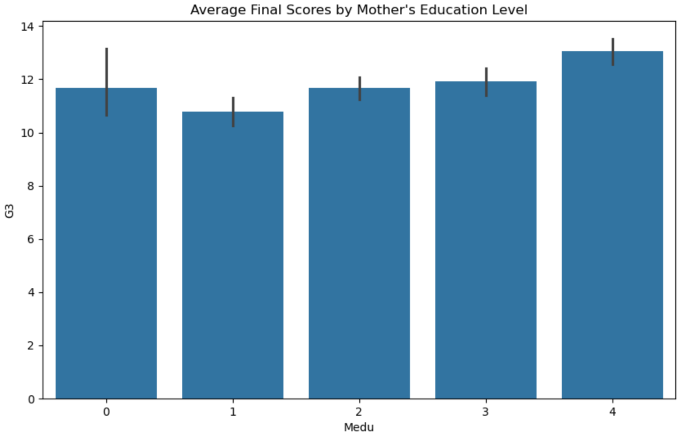

Figure 6: Mother’s Education Level vs Final Grade

Grade Category Distribution:

```
grade_cat = pd.cut(df[’G3’], bins=[0, 10, 15, 20], labels=[’Low’, ’Medium’, ’High’])
grade_cat.value_counts().plot.pie(autopct=’%1.1f%%’)
plt.title("Grade Category Distribution")
plt.ylabel(’’)
plt.savefig("pie_grade_category.png")
plt.show()
```

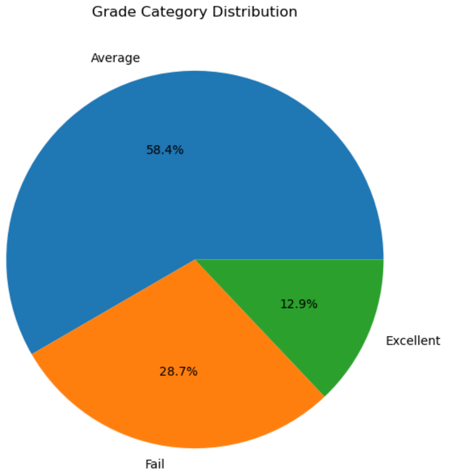

Figure 7: Distribution of Grade Categories

4 Task 4: Insights and Report Generation

- `G1` and `G2` are highly correlated with `G3`, indicating strong predictive value.
- Study time shows slight positive influence on grades, while failures and absences have negative effects.
- Students with more educated mothers tend to score better.

- Final grades are mostly concentrated in the 10–15 range (Medium).
- Features recommended for modeling: `G1`, `G2`, `failures`, `studytime`, and `absences`.