```python
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
from ucimlrepo import fetch_ucirepo

# Step 1: Load the dataset
dataset = fetch_ucirepo(id=320)
data_url = dataset['metadata']['data_url']

# Load the data from the URL
data = pd.read_csv(data_url)

# Extract variable names and data
variables = dataset['variables']
feature_names = variables['name'].tolist()
data.columns = feature_names

# Convert appropriate columns to numeric
for col in data.columns:
    data[col] = pd.to_numeric(data[col], errors='ignore')

# Identify the new target variable
target = 'G3'
```

```
C:\Users\Fujitsu\AppData\Local\Temp\ipykernel_2360\161631584.py:25:
FutureWarning: errors='ignore' is deprecated and will raise in a
future version. Use to_numeric without passing `errors` and catch
exceptions explicitly instead
  data[col] = pd.to_numeric(data[col], errors='ignore')
```

```python
# Initial Data Exploration (Basic Information)
print("Basic Information:")
print(data.info())
print("\nFirst few rows of the dataset:")
print(data.head())

print("\nSummary Statistics for Numerical Features:")
print(data.describe())

print("\nSummary Statistics for Categorical Features:")
print(data.describe(include=[object]))
```

```
Basic Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 33 columns):
```

```
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64
 3   address     649 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64
 7   Fedu        649 non-null    int64
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64
 13  studytime   649 non-null    int64
 14  failures    649 non-null    int64
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64
 24  freetime    649 non-null    int64
 25  goout       649 non-null    int64
 26  Dalc        649 non-null    int64
 27  Walc        649 non-null    int64
 28  health      649 non-null    int64
 29  absences    649 non-null    int64
 30  G1          649 non-null    int64
 31  G2          649 non-null    int64
 32  G3          649 non-null    int64
dtypes: int64(16), object(17)
memory usage: 167.4+ KB
None

First few rows of the dataset:
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob
Fjob  ...  \
0     GP   F   18       U     GT3       A     4     4  at_home
teacher  ...
1     GP   F   17       U     GT3       T     1     1  at_home
other  ...
2     GP   F   15       U     LE3       T     1     1  at_home
other  ...
3     GP   F   15       U     GT3       T     4     2   health
```

```
services  ...
4     GP   F   16       U     GT3       T     3     3    other
other  ...

  famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0      4        3      4     1     1      3        4   0  11  11
1      5        3      3     1     1      3        2   9  11  11
2      4        3      2     2     3      3        6  12  13  12
3      3        2      2     1     1      5        0  14  14  14
4      4        3      2     1     2      5        0  11  13  13

[5 rows x 33 columns]

Summary Statistics for Numerical Features:
              age        Medu        Fedu  traveltime   studytime
failures  \
count  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
std      1.218138    1.134552    1.099931    0.748660    0.829510
0.593235
min     15.000000    0.000000    0.000000    1.000000    1.000000
0.000000
25%     16.000000    2.000000    1.000000    1.000000    1.000000
0.000000
50%     17.000000    2.000000    2.000000    1.000000    2.000000
0.000000
75%     18.000000    4.000000    3.000000    2.000000    2.000000
0.000000
max     22.000000    4.000000    4.000000    4.000000    4.000000
3.000000

           famrel    freetime       goout        Dalc        Walc
health  \
count  649.000000  649.000000  649.000000  649.000000  649.000000
649.000000
mean     3.930663    3.180277    3.184900    1.502311    2.280431
3.536210
std      0.955717    1.051093    1.175766    0.924834    1.284380
1.446259
min      1.000000    1.000000    1.000000    1.000000    1.000000
1.000000
25%      4.000000    3.000000    2.000000    1.000000    1.000000
2.000000
50%      4.000000    3.000000    3.000000    1.000000    2.000000
4.000000
75%      5.000000    4.000000    4.000000    2.000000    3.000000
5.000000
max      5.000000    5.000000    5.000000    5.000000    5.000000
```

```
5.000000

         absences          G1          G2          G3
count  649.000000  649.000000  649.000000  649.000000
mean     3.659476   11.399076   11.570108   11.906009
std      4.640759    2.745265    2.913639    3.230656
min      0.000000    0.000000    0.000000    0.000000
25%      0.000000   10.000000   10.000000   10.000000
50%      2.000000   11.000000   11.000000   12.000000
75%      6.000000   13.000000   13.000000   14.000000
max     32.000000   19.000000   19.000000   19.000000

Summary Statistics for Categorical Features:
       school  sex address famsize Pstatus   Mjob   Fjob  reason
guardian  \
count     649  649     649     649     649    649    649     649
649
unique      2    2       2       2       2      5      5       4
3
top        GP    F       U     GT3       T  other  other  course
mother
freq      423  383     452     457     569    258    367     285
455

       schoolsup famsup paid activities nursery higher internet
romantic
count        649    649  649        649     649    649      649
649
unique         2      2    2          2       2      2        2
2
top           no    yes   no         no     yes    yes      yes
no
freq         581    398  610        334     521    580      498
410

# Step 3: Handling Missing Values
# Separate numerical and categorical features
numerical_features =
data.select_dtypes(include=[np.number]).columns.tolist()
categorical_features =
data.select_dtypes(include=[object]).columns.tolist()

# Impute missing values for numerical features
imputer_num = SimpleImputer(strategy='mean')
data[numerical_features] =
imputer_num.fit_transform(data[numerical_features])

# Explanation:
# Missing values can cause errors in data analysis and machine
```

```
learning models. Imputing with the mean is a common strategy for
numerical features to maintain data consistency.

# Step 4: Encoding Categorical Variables
# Convert categorical columns to numerical using one-hot encoding
data = pd.get_dummies(data, drop_first=True)

# Explanation:
# Categorical variables need to be converted to numerical form for
machine learning algorithms. One-hot encoding is a common method.

# Step 5: Outlier Removal
# Outlier removal using the IQR method for numerical features only
numeric_data = data.select_dtypes(include=[np.number])
Q1 = numeric_data.quantile(0.25)
Q3 = numeric_data.quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Only remove outliers for numerical columns
data = data[~((numeric_data < lower_bound) | (numeric_data >
upper_bound)).any(axis=1)]

# Explanation:
# Outliers can skew the results of data analysis and modeling. The IQR
method is used to identify and remove outliers from numerical features
only, ensuring a more robust analysis.

# Step 6: Normalizing Numerical Features
scaler = StandardScaler()
data[numerical_features] =
scaler.fit_transform(data[numerical_features])

# Explanation:
# Normalizing numerical features ensures that all features contribute
equally to the analysis and models, preventing features with larger
scales from dominating.
```

```
C:\Users\Fujitsu\AppData\Local\Temp\ipykernel_2360\265086805.py:3:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  data[numerical_features] =
scaler.fit_transform(data[numerical_features])
```

```python
# Step 7: Splitting the Dataset into Train and Test Sets
X = data.drop(columns=[target])
y = data[target]
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Explanation:
# Splitting the data into training and testing sets allows for the
evaluation of model performance on unseen data, helping to prevent
overfitting.

# Display basic information about the processed data
print("Basic Information after Preprocessing:")
print(data.info())

print("\nFirst few rows of the processed dataset:")
print(data.head())

print("\nSummary Statistics of the processed dataset:")
print(data.describe())

# Save the cleaned dataset for further analysis
data.to_csv('cleaned_student_data2.csv', index=False)
print("Data preprocessing and cleaning complete.")
```

```
Basic Information after Preprocessing:
<class 'pandas.core.frame.DataFrame'>
Index: 393 entries, 1 to 648
Data columns (total 42 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   age                393 non-null    float64
 1   Medu               393 non-null    float64
 2   Fedu               393 non-null    float64
 3   traveltime         393 non-null    float64
 4   studytime          393 non-null    float64
 5   failures           393 non-null    float64
 6   famrel             393 non-null    float64
 7   freetime           393 non-null    float64
 8   goout              393 non-null    float64
 9   Dalc               393 non-null    float64
 10  Walc               393 non-null    float64
 11  health             393 non-null    float64
 12  absences           393 non-null    float64
 13  G1                 393 non-null    float64
 14  G2                 393 non-null    float64
 15  G3                 393 non-null    float64
 16  school_MS          393 non-null    bool
 17  sex_M              393 non-null    bool
 18  address_U          393 non-null    bool
```

```
 19  famsize_LE3        393 non-null    bool
 20  Pstatus_T          393 non-null    bool
 21  Mjob_health        393 non-null    bool
 22  Mjob_other         393 non-null    bool
 23  Mjob_services      393 non-null    bool
 24  Mjob_teacher       393 non-null    bool
 25  Fjob_health        393 non-null    bool
 26  Fjob_other         393 non-null    bool
 27  Fjob_services      393 non-null    bool
 28  Fjob_teacher       393 non-null    bool
 29  reason_home        393 non-null    bool
 30  reason_other       393 non-null    bool
 31  reason_reputation  393 non-null    bool
 32  guardian_mother    393 non-null    bool
 33  guardian_other     393 non-null    bool
 34  schoolsup_yes      393 non-null    bool
 35  famsup_yes         393 non-null    bool
 36  paid_yes           393 non-null    bool
 37  activities_yes     393 non-null    bool
 38  nursery_yes        393 non-null    bool
 39  higher_yes         393 non-null    bool
 40  internet_yes       393 non-null    bool
 41  romantic_yes       393 non-null    bool
dtypes: bool(26), float64(16)
memory usage: 62.2 KB
None

First few rows of the processed dataset:
        age      Medu      Fedu  traveltime  studytime  failures
famrel  \
1  0.433258 -1.399119 -1.302020   -0.712685   0.142300       0.0
1.320544
2 -1.387817 -1.399119 -1.302020   -0.712685   0.142300       0.0 -
0.224021
3 -1.387817  1.223395 -0.397967   -0.712685   1.576243       0.0 -
1.768586
4 -0.477279  0.349224  0.506086   -0.712685   0.142300       0.0 -
0.224021
5 -0.477279  1.223395  0.506086   -0.712685   0.142300       0.0
1.320544

   freetime     goout      Dalc  ...  guardian_mother  guardian_other
\
1 -0.328816 -0.144317 -0.516951  ...            False           False

2 -0.328816 -1.044581  1.329973  ...             True           False

3 -1.472399 -1.044581 -0.516951  ...             True           False

4 -0.328816 -1.044581 -0.516951  ...            False           False
```

```
5  0.814766 -1.044581 -0.516951  ...             True           False


   schoolsup_yes  famsup_yes  paid_yes  activities_yes  nursery_yes  \
1          False        True     False           False        False
2           True       False     False           False         True
3          False        True     False            True         True
4          False        True     False           False         True
5          False        True     False            True         True

   higher_yes  internet_yes  romantic_yes
1        True          True         False
2        True          True         False
3        True          True          True
4        True         False         False
5        True          True         False

[5 rows x 42 columns]

Summary Statistics of the processed dataset:
                age          Medu          Fedu    traveltime
studytime  \
count  3.930000e+02  3.930000e+02  3.930000e+02  3.930000e+02
3.930000e+02
mean   6.147189e-16  2.169596e-16 -1.807997e-17 -7.231987e-17
1.378598e-16
std    1.001275e+00  1.001275e+00  1.001275e+00  1.001275e+00
1.001275e+00
min   -1.387817e+00 -2.273290e+00 -2.206073e+00 -7.126850e-01 -
1.291644e+00
25%   -4.772791e-01 -5.249476e-01 -3.979673e-01 -7.126850e-01 -
1.291644e+00
50%   -4.772791e-01  3.492236e-01 -3.979673e-01 -7.126850e-01
1.422998e-01
75%    4.332582e-01  1.223395e+00  5.060856e-01  8.787082e-01
1.422998e-01
max    3.164870e+00  1.223395e+00  1.410138e+00  2.470101e+00
1.576243e+00

       failures        famrel      freetime         goout
Dalc  \
count     393.0  3.930000e+02  3.930000e+02  3.930000e+02
3.930000e+02
mean        0.0 -5.514390e-16  1.717597e-16 -3.615994e-17  9.039984e-
17
std         0.0  1.001275e+00  1.001275e+00  1.001275e+00
1.001275e+00
min         0.0 -1.768586e+00 -1.472399e+00 -1.944845e+00 -5.169506e-
01
```

```
25%          0.0 -2.240209e-01 -3.288164e-01 -1.044581e+00 -5.169506e-
01
50%          0.0 -2.240209e-01 -3.288164e-01 -1.443171e-01 -5.169506e-
01
75%          0.0  1.320544e+00  8.147663e-01  7.559468e-01 -5.169506e-
01
max          0.0  1.320544e+00  1.958349e+00  1.656211e+00
3.176897e+00

               Walc        health      absences            G1
G2  \
count  3.930000e+02  3.930000e+02  3.930000e+02  3.930000e+02
3.930000e+02
mean   1.220398e-16 -1.807997e-17  2.259996e-17 -4.519992e-17
9.491983e-17
std    1.001275e+00  1.001275e+00  1.001275e+00  1.001275e+00
1.001275e+00
min   -9.843020e-01 -1.814934e+00 -8.694999e-01 -2.582509e+00 -
2.317264e+00
25%   -9.843020e-01 -4.108610e-01 -8.694999e-01 -8.386499e-01 -
5.333565e-01
50%   -1.011269e-01  2.911754e-01 -2.419270e-01  3.327976e-02 -
8.737969e-02
75%    7.820481e-01  9.932119e-01  3.856460e-01  9.052094e-01
8.045740e-01
max    2.548398e+00  9.932119e-01  3.523511e+00  2.213104e+00
2.142505e+00

                 G3
count  3.930000e+02
mean  -2.711995e-17
std    1.001275e+00
min   -2.431948e+00
25%   -7.198915e-01
50%    1.361368e-01
75%    5.641510e-01
max    2.276208e+00
Data preprocessing and cleaning complete.
```