DATA ANALYST

Intenship Task 4

DESCRIPTION

Cleaned and prepared a Student Performance dataset using Python (Pandas) in Google Colab. Performed data inspection, verified missing values and duplicates, standardized column names, and created new features such as total score and performance level. Exported the final cleaned dataset for analysis.

PREPARED BY

Reema Safrin M
(22-01-2026)

MY WORK

AData Cleaning and Preparation

- Imported the student performance dataset using Pandas
- Inspected dataset structure using info()
- Checked for missing values and found none
- Checked for duplicate records and found none
- Standardized column naming for clarity
- Created new features such as total_score and performance_level
- Saved the cleaned dataset for further analysis

DATASET

students_perfomence_dataset

CLEANED_DATASET

students_perfomence_dataset

MY GOOGLE COLAB WORK

```
from google.colab import files
files.upload()
```

Choose Files student-per…-dataset.csv

student-perfomance-dataset.csv(text/csv) - 69361 bytes, last modified: 1/22/2026 - 100% done

```
Saving student-perfomance-dataset.csv to student-perfomance-dataset(1).csv
{'student-perfomance-dataset (1).csv':
b'school,sex,age,address,family_size,parent_status,mother_education,father_education
```

```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv("student-perfomance-dataset.csv")
df.head()
```

	school	sex	age	address	family_size	parent_status	mother_education	father_ed
	0GPF18	**1**GPF17		U	GT3	A	4	
	2GPF15	**3**GPF15		U	GT3	T	1	
	4GPF16			U	LE3	T	1	
				U	GT3	T	4	
				U	GT3	T	3	

5 rows × 33 columns

```
df.info()
```

```
<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 family_size             649 non-null    object
5 parent_status           649 non-null    object
6 mother_education        649 non-null    int64
7 father_education        649 non-null    int64
8 mother_job              649 non-null    object
9 father_job              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 school_support         649 non-null    object
16 family_support         649 non-null    object
17 paid                   649 non-null    object
18 extra_activities       649 non-null    object
19 nursery                649 non-null    object
20 higher_education       649 non-null    object
```

```
      21     internet_access           22   object
romantic_relationship  649  non-null  649   object
non-null                                    int64
649 non-null                                int64
649  non-null  649  non-null  649  non-null  int64
649 non-null                                int64
649 non-null                                int64
649 non-null 649 non-null 649 non-null      int64
                                            int64
                                            int64
                                            int64
                                            int64

dtypes: int64(16), object(17)
memory usage: 167.4+ KB
```

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

	0
school	0
sex	0
age	0
address	0
family_size	0
parent_status	0
mother_education	0
father_education	0
mother_job	0
father_job	0
reason	0
guardian	0
traveltime	0
studytime	0
failures	0
school_support	0
family_support	0
paid	0
extra_activities	0
nursery	0
higher_education	0
internet_access	0
romantic_relationship	0
family_relationship	0
freetime	0
going_out	0
workday_alcohol	0
weekend_alcohol	0
health	0
absences	0
grade_1	0
grade_2	0
final_grade	0

dtype: int64

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

```
np.int64(0)
```

```
df.drop_duplicates(inplace=True)
```

```
df['total_score'] = df['grade_1'] + df['grade_2'] + df['final_grade']
```

```
df['performance_level'] = pd.cut(
    df['total_score'],
    bins=[0, 150, 210, 300],
    labels=['Low', 'Average', 'High']
)
```

```
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 35 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  family_size           649 non-null    object
5  parent_status         649 non-null    object
6  mother_education      649 non-null    int64 
7  father_education      649 non-null    int64 
8  mother_job            649 non-null    object
9  father_job            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 school_support        649 non-null    object
16 family_support        649 non-null    object
```