DATA ANALYST

Intenship Task 10

DESCRIPTION

The objective of this task was to perform Exploratory Data Analysis (EDA) on the Student Performance dataset using Python. It involved generating summary statistics, analyzing missing values, visualizing data distributions, detecting outliers using the IQR method, handling extreme values, and computing correlations to understand relationships among variables.

PREPARED BY

Reema Safrin M
(30-01-2026)

MY WORK

In this task, I loaded the cleaned Student Performance dataset into Google Colab using Python libraries such as pandas, numpy, and matplotlib. I explored the dataset structure using shape, info, and head functions and generated descriptive statistics to understand data distribution. I calculated missing value percentages to ensure data quality and visualized numeric features using histograms and boxplots. Outliers were detected using the Interquartile Range (IQR) method and handled through capping to minimize extreme impact. A correlation matrix was created to identify relationships among academic performance variables. Finally, the cleaned dataset and EDA findings were exported successfully.

DATASET USED

cleaned_student_performance

FINAL DATASET

final_student_performance

MY GOOGLE COLAB DATA

Student Performance – EDA & Outlier Detection Report

1. Dataset Overview

Total Rows: 649

Total Columns: 36

2. Missing Values

	Missing Count
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
total_score	0

	Missing Count
performance_level	0
Outlier_Flag	0

3. Descriptive Statistics

	age	mother_education	father_education	traveltime	studytime	failures	family_relationship	freetime	going_out
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.0	649.000000	649.000000	649.000000
mean	16.742681	2.514638	2.306626	1.556240	1.903698	0.0	4.003852	3.214946	3.18[illegible]900
std	1.212097	1.134552	1.099931	0.711672	0.767523	0.0	0.781480	0.984655	1.17[illegible]766
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.0	2.500000	1.500000	1.00[illegible]000
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.0	4.000000	3.000000	2.00[illegible]000
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.0	4.000000	3.000000	3.00[illegible]000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.0	5.000000	4.000000	4.00[illegible]000
max	21.000000	4.000000	4.000000	3.500000	3.500000	0.0	5.000000	5.000000	5.00[illegible]000

4. Sample Data

	school	sex	age	address	family_size	parent_status	mother_education	father_education	mother_job	father_job	reason	gu
0	GP	F	18.0	U	GT3	A	4.0	4.0	at_home	teacher	course	m
1	GP	F	17.0	U	GT3	T	1.0	1.0	at_home	other	course	f
2	GP	F	15.0	U	LE3	T	1.0	1.0	at_home	other	other	m
3	GP	F	15.0	U	GT3	T	4.0	2.0	health	services	home	m
4	GP	F	16.0	U	GT3	T	3.0	3.0	other	other	home	f
5	GP	M	16.0	U	LE3	T	4.0	3.0	services	other	reputatio	m
6	GP	M	16.0	U	LE3	T	2.0	2.0	other	other	home	m
7	GP	F	17.0	U	GT3	A	4.0	4.0	other	teacher	home	m
8	GP	M	15.0	U	LE3	A	3.0	2.0	services	other	home	m
9	GP	M	15.0	U	GT3	T	3.0	4.0	other	other	home	m

5. Outlier Detection Method

The IQR (Interquartile Range) method was used to detect outliers in numeric columns.

Outliers were handled using the capping method to limit extreme values.

6. Correlation Matrix

	age	mother_education	father_education	traveltime	studytime	failures	family_relationship	freetime	go
age	1.000000	-0.108914	-0.120141	0.043637	0.004010	NaN	-0.022575	-0.00660[illegible]	0.
mother_education	-0.108914	1.000000	0.647477	-0.266209	0.098649	NaN	0.026479	-0.02526[illegible]	0.
father_education	-0.120141	0.647477	1.000000	-0.210335	0.060623	NaN	0.018372	0.00245[illegible]	0.
traveltime	0.043637	-0.266209	-0.210335	1.000000	-0.078358	NaN	-0.004552	0.00419[illegible]	0.

	age	mother_education	father_education	traveltime	studytime	failures	family_relationship	freetime	go
studytime	0.004010	0.098649	0.060623	-0.078358	1.000000	NaN	0.019916	-0.08028	-0
failures	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
family_relationship	-0.022575	0.026479	0.018372	-0.004552	0.019916	NaN	1.000000	0.13880	0.
freetime	-0.006600	-0.025268	0.002459	0.004194	-0.080282	NaN	0.138806	1.00000	0.
going_out	0.111401	0.009536	0.027690	0.053947	-0.078567	NaN	0.087399	0.35084	1.
workday_alcohol	0.116683	-0.004071	0.002706	0.092258	-0.161097	NaN	-0.075692	0.11805	0.
weekend_alcohol	0.084092	-0.019766	0.038445	0.052784	-0.223818	NaN	-0.099478	0.12309	0.
health	-0.006561	0.004614	0.044910	-0.051085	-0.055591	NaN	0.095796	0.08040	-0
absences	0.154296	-0.017426	0.027644	0.002062	-0.118162	NaN	-0.101868	-0.01981	0.
grade_1	-0.172384	0.271380	0.228412	-0.158937	0.267515	NaN	0.031858	-0.10569	-0
grade_2	-0.096108	0.272815	0.229548	-0.158567	0.254284	NaN	0.070946	-0.11444	-0
final_grade	-0.088829	0.259756	0.217141	-0.132426	0.267273	NaN	0.048406	-0.13403	-0
total_score	-0.124382	0.271909	0.229292	-0.154322	0.270782	NaN	0.050496	-0.12440	-0

7. Key Observations

- Most numeric features show normal to slightly skewed distributions.
- Outliers were present and controlled using IQR capping.
- Strong correlations exist between subject performance scores.
- Dataset had minimal missing values after cleaning.