Initial Summary Report

==========

First 5 Rows:

	StudentID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	\
0	1001	17	1	0	2	19.833723	
1	1002	18	0	0	1	15.408756	
2	1003	15	0	2	3	4.210570	
3	1004	17	1	0	3	10.028829	
4	1005	17	1	0	2	4.672495	

	Absences	Tutoring	ParentalSupport	Extracurricular	Sports	Music	\
0	7	1	2	0	0	1	
1	0	0	1	0	0	0	
2	26	0	2	0	0	0	
3	14	0	3	1	0	0	
4	17	1	3	0	0	0	

	Volunteering	GPA	GradeClass
0	0	2.929196	2.0
1	0	3.042915	1.0
2	0	0.112602	4.0
3	0	2.054218	3.0
4	0	1.288061	4.0

Data Info:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2392 entries, 0 to 2391

Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
---	------	--------------	-----

```
 0  StudentID        2392 non-null   int64
 1  Age              2392 non-null   int64
 2  Gender           2392 non-null   int64
 3  Ethnicity        2392 non-null   int64
 4  ParentalEducation  2392 non-null   int64
 5  StudyTimeWeekly    2392 non-null   float64
 6  Absences         2392 non-null   int64
 7  Tutoring         2392 non-null   int64
 8  ParentalSupport    2392 non-null   int64
 9  Extracurricular    2392 non-null   int64
 10  Sports          2392 non-null   int64
 11  Music           2392 non-null   int64
 12  Volunteering      2392 non-null   int64
 13  GPA             2392 non-null   float64
 14  GradeClass       2392 non-null   float64
dtypes: float64(3), int64(12)
memory usage: 280.4 KB
```

Basic Statistics:

```
       StudentID        Age      Gender    Ethnicity  ParentalEducation  \
count  2392.000000  2392.000000  2392.000000  2392.000000      2392.000000
mean   2196.500000    16.468645     0.510870     0.877508         1.746237
std     690.655244     1.123798     0.499986     1.028476         1.000411
min    1001.000000    15.000000     0.000000     0.000000         0.000000
25%    1598.750000    15.000000     0.000000     0.000000         1.000000
50%    2196.500000    16.000000     1.000000     0.000000         2.000000
75%    2794.250000    17.000000     1.000000     2.000000         2.000000
max    3392.000000    18.000000     1.000000     3.000000         4.000000

       StudyTimeWeekly    Absences     Tutoring  ParentalSupport  \
```

count	2392.000000	2392.000000	2392.000000	2392.000000
mean	9.771992	14.541388	0.301421	2.122074
std	5.652774	8.467417	0.458971	1.122813
min	0.001057	0.000000	0.000000	0.000000
25%	5.043079	7.000000	0.000000	1.000000
50%	9.705363	15.000000	0.000000	2.000000
75%	14.408410	22.000000	1.000000	3.000000
max	19.978094	29.000000	1.000000	4.000000

	Extracurricular	Sports	Music	Volunteering	GPA \
count	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000
mean	0.383361	0.303512	0.196906	0.157191	1.906186
std	0.486307	0.459870	0.397744	0.364057	0.915156
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	1.174803
50%	0.000000	0.000000	0.000000	0.000000	1.893393
75%	1.000000	1.000000	0.000000	0.000000	2.622216
max	1.000000	1.000000	1.000000	1.000000	4.000000

	GradeClass
count	2392.000000
mean	2.983696
std	1.233908
min	0.000000
25%	2.000000
50%	4.000000
75%	4.000000
max	4.000000

Missing Values Check:

StudentID	0

Age 0

Gender 0

Ethnicity 0

ParentalEducation 0

StudyTimeWeekly 0

Absences 0

Tutoring 0

ParentalSupport 0

Extracurricular 0

Sports 0

Music 0

Volunteering 0

GPA 0

GradeClass 0

dtype: int64

Missing Values After:

StudentID 0

Age 0

Gender 0

Ethnicity 0

ParentalEducation 0

StudyTimeWeekly 0

Absences 0

Tutoring 0

ParentalSupport 0

Extracurricular 0

Sports 0

Music 0

Volunteering 0

GPA 0

GradeClass 0

dtype: int64

Outlier Treatment:

Applied IQR-based capping (multiplier=1.5) to StudyTimeWeekly, Absences, GPA.

Visualizations saved as plots/outliers_before.png and plots/outliers_after.png.

Distributions:

Numeric: StudyTimeWeekly, Absences, GPA saved in plots/distributions.png.

Categorical: GradeClass, Gender, Ethnicity, ParentalSupport saved in plots/categorical_distributions.png.

Conclusion:

Dataset cleaned and saved as cleaned_student_performance_data.csv.

All missing values handled and outliers capped.

Visualizations and summaries provided for further analysis.