In [98]:
```
import pandas as pd
# Load the dataset, treating '?' as missing values
df = pd.read_csv('Project Step-2 Data Collection Group 12 - Sheet1.csv', na_values=['?'])
df
```

Out[98]:

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offer Electives?	Offers Sports?	Crime Related Data	...	Math Score	English Score	Racial% White	Racial% Black	Racial% Hispanic	Racial% Asian	R
0	Boulder Creek High School	Anthem	85086	565th	23rd	Yes	Yes	NaN	Yes	21	...	43	46	77.7	2	13.9	1.5	
1	Great Hearts Academics	Anthem	85086	7,645th	152nd	No	No	NaN	Yes	1	...	86	78	69.3	1.3	15.3	6.1	
2	Agua Fria High School	Avondale	85323	2,423rd	86th	Yes	Yes	NaN	Yes	48	...	25	38	13.4	11.5	69.8	1.7	
3	Arizona Agribusiness & Equine Center - Estrella	Avondale	85007	11,110th	201st	No	Yes	NaN	NaN	0	...	37	57	30.4	4.4	52.6	6.2	
4	E-Institute at Avondale	Avondale	85323	21,288th	434th	NaN	NaN	NaN	NaN	0	...	NaN	NaN	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
195	West-Mec - Cortez High School	Phoenix	85051	25,823	609	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
196	West-Mec - Greenway High School	Phoenix	85053	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
197	West-Mec - Moon Valley High School	Phoenix	85029	25,823	609	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
198	Tumbleweed Transitional Learning Center	Phoenix	85013	22,127	462	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
199	System Phoenix	Phoenix	85006	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	

200 rows × 24 columns

In [99]:
```
# Step 1: Calculate the number of missing values per row
missing_values_per_row = df.isna().sum(axis=1)
print("Missing values per row:\n", missing_values_per_row)
```

```
Missing values per row:
 0       1
1       5
2       1
3       6
4      17
      ..
195    19
196    21
197    19
198    19
199    21
Length: 200, dtype: int64
```

In [100…
```
df['Student Teacher Ratio']
```

Out[100]:
```
0      25:1
1       NaN
2      17:1
3       NaN
4       NaN
       ...
195     NaN
196     NaN
197     NaN
198     NaN
199     NaN
Name: Student Teacher Ratio, Length: 200, dtype: object
```

In [101…
```
# Step 1: Calculate the number of missing values per column
missing_values_per_column = df.isna().sum()
print("\nMissing values per column:\n", missing_values_per_column)
```

```
Missing values per column:
 School Name                          0
City                                  0
Zip Code                              0
National Rank                        14
Arizona Rank                         14
AP Classes                           41
Dual Enrollments                     39
Offer Electives?                    140
Offers Sports?                       52
Crime Related Data                   27
Teaching and Educational Method      38
Mental Health Services              111
2022N/A2023 Student Enrollment       25
Student Teacher Ratio               100
Math Score                           43
English Score                        44
Racial% White                        30
Racial% Black                        33
Racial% Hispanic                     30
Racial% Asian                        42
Racial% Other                        32
Lunch % Free                         93
Lunch% Reduced                       93
Lunch%Paid                           94
dtype: int64
```

In [102…
```
# Step 2: Experiment with removing columns with the maximum number of missing values
max_missing_columns = missing_values_per_column.idxmax()
print("\nColumn with the maximum number of missing values:", max_missing_columns)
```

```
Column with the maximum number of missing values: Offer Electives?
```

In [103…
```
# Drop the column with the maximum number of missing values
df_cleaned = df.drop(columns=max_missing_columns)
```

In [104…
```
# Step 3: Final data cleaning procedure
# Drop rows with any missing values
df_cleaned = df_cleaned.dropna()
```

In [105…
```
# Convert "Student Teacher Ratio" column to string/object data type
df_cleaned['Student Teacher Ratio'] = df_cleaned['Student Teacher Ratio'].astype(str)
```

In [106…
```
# Print the cleaned dataset
df_cleaned
```

Out[106]:

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offers Sports?	Crime Related Data	Teaching and Educational Method	...	Math Score	English Score	Racia Wh
0	Boulder Creek High School	Anthem	85086	565th	23rd	Yes	Yes	Yes	21	Hybrid	...	43	46	7
2	Agua Fria High School	Avondale	85323	2,423rd	86th	Yes	Yes	Yes	48	Hybrid	...	25	38	1
12	Buckeye Union High School	Buckeye	85326	3,772nd	109th	Yes	Yes	Yes	36	In-class	...	21	30	2
15	Youngker High School	Buckeye	85326	3,378th	106th	Yes	Yes	Yes	46	In-class	...	21	30	2
20	Blueprint High School	Chandler	85225	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	...	Permanently Closed	Permanently Closed	Permanen Clo
46	Deer Valley High School	Glendale	85308	2,141st	76th	Yes	Yes	Yes	38	Hybrid	...	43	46	5
48	Desert Sky Middle School	Glendale	85308	Unranked	Unknown	No	No	Yes	31	In-class	...	49	52	5
50	Apollo High School	Glendale	85302	1,489	49th	Yes	Yes	Yes	47	Hybrid	...	18	30	
51	Cactus High School	Glendale	85302	4,643	119th	Yes	Yes	Yes	3	Hybrid	...	30	33	5
56	Desert Edge High School	Goodyear	85338	2,302	81st	Yes	Yes	Yes	39	Hybrid	...	25	38	2
57	Estrella Foothills High School	Goodyear	85338	5,773rd	133rd	Yes	Yes	Yes	0	In-class	...	21	30	4
60	Desert Edge High School	Goodyear	85338	2,302	81st	Yes	Yes	Yes	39	Hybrid	...	25	38	2
61	Estrella Foothills High School	Goodyear	85338	5,773	133	Yes	Yes	Yes	12	In-class	...	21	30	4
66	Cesar Chavez High School	Laveen	85339	478	18th	Yes	Yes	Yes	26	Hybrid	...	24	24	
114	Desert Mountain High School	Scottsdale	85259	35	1009	Yes	Yes	Yes	26	Hybrid	...	49	47	74.8
121	Imagine Prep Surprise	Surprise	85379	265	14874	No	Yes	Yes	0	Hybrid	...	17	27	53.9
124	Valley Vista High School	Surprise	85374	27	608	Yes	No	Yes	104	Hybrid	...	25	32	35.3
136	Sierra Linda High School	Tolleson	85353	62	1898	Yes	Yes	Yes	52	Hybrid	...	10	14	3.3
163	Maryvale High School	Phoenix	85033	273	14	Yes	Yes	Yes	24	In-class	...	24	24	
165	Moon Valley High School	Phoenix	85029	3,474	107	Yes	Yes	Yes	36	In-class	...	18	30	3
167	North Canyon High School	Phoenix	85024	1,907	64	Yes	Yes	Yes	50	Hybrid	...	33	40	2
168	North High School	Phoenix	85014	573	24	Yes	Yes	Yes	27	In-class	...	24	24	
173	Paradise Valley High	Phoenix	85032	2,267	79	Yes	Yes	Yes	21	Hybrid	...	33	40	3

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offers Sports?	Crime Related Data	Teaching and Educational Method	...	Math Score	English Score	Raci Wh
	School													
176	Pinnacle High School	Phoenix	85050	696	28	Yes	Yes	Yes	18	In-class	...	33	40	7
177	Roadrunner School	Phoenix	85028	17,197	312	No	No	No	8	In-class	...	33	40	
178	Shadow Mountain High School	Phoenix	85028	3,930	111	Yes	Yes	Yes	50	Hybrid	...	33	40	
190	Vista Peak	Phoenix	85027	22,127	462	No	No	No	2	In-class	...	43	46	2

27 rows × 22 columns

In [107…
```
df_cleaned['Student Teacher Ratio']
```

Out[107]:
```
0                    25:1
2                    17:1
12                   19:1
15                   23:1
20     Permanently Closed
46                   10:1
48                   22:1
50                   24:1
51                   19:1
56                   21:1
57                   25:1
60                   21:1
61                   25:1
66                   20:1
114                23.1:1
121                N/A1:1
124                26.8:1
136                24.3:1
163                  21:1
165                  20:1
167                  23:1
168                  18:1
173                  18:1
176                  24:1
177                   4:1
178                  15:1
190                   4:1
Name: Student Teacher Ratio, dtype: object
```

In [108…
```
# Step 4: Detect and remove duplicate records
# Group the data by school name, city, and zip code
grouped = df_cleaned.groupby(['School Name', 'City', 'Zip Code'])

# Check for multiple records for each group
duplicate_groups = grouped.filter(lambda x: len(x) > 1)

# Print duplicate records
print("\nDuplicate records:")
duplicate_groups
```

```
Duplicate records:
```

Out[108]:

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offers Sports?	Crime Related Data	Teaching and Educational Method	...	Math Score	English Score	Racial% White	Racial% Black	Racial% Hispanic	Racial% Asian	Rac O
56	Desert Edge High School	Goodyear	85338	2,302	81st	Yes	Yes	Yes	39	Hybrid	...	25	38	22.2	12.7	57.2	2.7	
57	Estrella Foothills High School	Goodyear	85338	5,773rd	133rd	Yes	Yes	Yes	0	In-class	...	21	30	45.6	4.5	43.1	1.3	
60	Desert Edge High School	Goodyear	85338	2,302	81st	Yes	Yes	Yes	39	Hybrid	...	25	38	22.2	12.7	57.2	2.7	
61	Estrella Foothills High School	Goodyear	85338	5,773	133	Yes	Yes	Yes	12	In-class	...	21	30	45.6	4.5	43.1	1.3	

4 rows × 23 columns

In [109…

```
# Combine duplicate records into one record
cleaned_df = df_cleaned.groupby(['School Name', 'City', 'Zip Code'], as_index=False).first()

# Print cleaned dataset
print("\nCleaned dataset after removing duplicates:")
cleaned_df
```

```
Cleaned dataset after removing duplicates:
```

Out[109]:

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offers Sports?	Crime Related Data	Teaching and Educational Method	...	Math Score	English Score	Racial Whi
0	Agua Fria High School	Avondale	85323	2,423rd	86th	Yes	Yes	Yes	48	Hybrid	...	25	38	13
1	Apollo High School	Glendale	85302	1,489	49th	Yes	Yes	Yes	47	Hybrid	...	18	30	
2	Blueprint High School	Chandler	85225	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	Permanently Closed	...	Permanently Closed	Permanently Closed	Permanent Clos
3	Boulder Creek High School	Anthem	85086	565th	23rd	Yes	Yes	Yes	21	Hybrid	...	43	46	77
4	Buckeye Union High School	Buckeye	85326	3,772nd	109th	Yes	Yes	Yes	36	In-class	...	21	30	23
5	Cactus High School	Glendale	85302	4,643	119th	Yes	Yes	Yes	3	Hybrid	...	30	33	52
6	Cesar Chavez High School	Laveen	85339	478	18th	Yes	Yes	Yes	26	Hybrid	...	24	24	4
7	Deer Valley High School	Glendale	85308	2,141st	76th	Yes	Yes	Yes	38	Hybrid	...	43	46	55
8	Desert Edge High School	Goodyear	85338	2,302	81st	Yes	Yes	Yes	39	Hybrid	...	25	38	22
9	Desert Mountain High School	Scottsdale	85259	35	1009	Yes	Yes	Yes	26	Hybrid	...	49	47	74.80
10	Desert Sky Middle School	Glendale	85308	Unranked	Unknown	No	No	Yes	31	In-class	...	49	52	51
11	Estrella Foothills High School	Goodyear	85338	5,773rd	133rd	Yes	Yes	Yes	0	In-class	...	21	30	45
12	Imagine Prep Surprise	Surprise	85379	265	14874	No	Yes	Yes	0	Hybrid	...	17	27	53.90
13	Maryvale High School	Phoenix	85033	273	14	Yes	Yes	Yes	24	In-class	...	24	24	2
14	Moon Valley High School	Phoenix	85029	3,474	107	Yes	Yes	Yes	36	In-class	...	18	30	30
15	North Canyon High School	Phoenix	85024	1,907	64	Yes	Yes	Yes	50	Hybrid	...	33	40	22
16	North High School	Phoenix	85014	573	24	Yes	Yes	Yes	27	In-class	...	24	24	3
17	Paradise Valley High School	Phoenix	85032	2,267	79	Yes	Yes	Yes	21	Hybrid	...	33	40	39
18	Pinnacle High School	Phoenix	85050	696	28	Yes	Yes	Yes	18	In-class	...	33	40	74
19	Roadrunner School	Phoenix	85028	17,197	312	No	No	No	8	In-class	...	33	40	
20	Shadow Mountain High School	Phoenix	85028	3,930	111	Yes	Yes	Yes	50	Hybrid	...	33	40	
21	Sierra Linda High School	Tolleson	85353	62	1898	Yes	Yes	Yes	52	Hybrid	...	10	14	3.30

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offers Sports?	Crime Related Data	Teaching and Educational Method	...	Math Score	English Score	Racial Whi
22	Valley Vista High School	Surprise	85374	27	608	Yes	No	Yes	104	Hybrid	...	25	32	35.30
23	Vista Peak	Phoenix	85027	22,127	462	No	No	No	2	In-class	...	43	46	21
24	Youngker High School	Buckeye	85326	3,378th	106th	Yes	Yes	Yes	46	In-class	...	21	30	22

25 rows × 22 columns

In [110…
```
df_cleaned['Student Teacher Ratio']
```

Out[110]:
```
0                    25:1
2                    17:1
12                   19:1
15                   23:1
20     Permanently Closed
46                   10:1
48                   22:1
50                   24:1
51                   19:1
56                   21:1
57                   25:1
60                   21:1
61                   25:1
66                   20:1
114                23.1:1
121                N/A1:1
124                26.8:1
136                24.3:1
163                  21:1
165                  20:1
167                  23:1
168                  18:1
173                  18:1
176                  24:1
177                   4:1
178                  15:1
190                   4:1
Name: Student Teacher Ratio, dtype: object
```

In [111…
```
# Step 5: Transform the data

# Define a function to remove special characters from numeric values
def remove_special_chars(value):
    if isinstance(value, str):
        return ''.join(char for char in value if char.isdigit() or char == '.')
    return value

# Apply the function to relevant numeric columns
numeric_columns = ['National Rank', 'Arizona Rank', 'Crime Related Data',
                   '2022N/A2023 Student Enrollment',
                   'Math Score', 'English Score', 'Racial% White', 'Racial% Black', 'Racial% Hispanic',
                   'Racial% Asian', 'Racial% Other', 'Lunch % Free', 'Lunch% Reduced', 'Lunch%Paid']
df_cleaned[numeric_columns] = df_cleaned[numeric_columns].applymap(remove_special_chars)

df_cleaned
```

Out[111]:

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offers Sports?	Crime Related Data	Teaching and Educational Method	...	Math Score	English Score	Racial% White	Racial% Black	Racial% Hispanic
0	Boulder Creek High School	Anthem	85086	565	23	Yes	Yes	Yes	21	Hybrid	...	43	46	77.7	2	13.9
2	Agua Fria High School	Avondale	85323	2423	86	Yes	Yes	Yes	48	Hybrid	...	25	38	13.4	11.5	69.8
12	Buckeye Union High School	Buckeye	85326	3772	109	Yes	Yes	Yes	36	In-class	...	21	30	23.9	7.5	64.4
15	Youngker High School	Buckeye	85326	3378	106	Yes	Yes	Yes	46	In-class	...	21	30	22.8	8.8	64.1
20	Blueprint High School	Chandler	85225			Permanently Closed	Permanently Closed	Permanently Closed		Permanently Closed	...					
46	Deer Valley High School	Glendale	85308	2141	76	Yes	Yes	Yes	38	Hybrid	...	43	46	55.2	5.9	30.5
48	Desert Sky Middle School	Glendale	85308			No	No	Yes	31	In-class	...	49	52	51.3	3.3	36.1
50	Apollo High School	Glendale	85302	1489	49	Yes	Yes	Yes	47	Hybrid	...	18	30	14	8.3	70.2
51	Cactus High School	Glendale	85302	4643	119	Yes	Yes	Yes	3	Hybrid	...	30	33	52.5	5.1	33.2
56	Desert Edge High School	Goodyear	85338	2302	81	Yes	Yes	Yes	39	Hybrid	...	25	38	22.2	12.7	57.2
57	Estrella Foothills High School	Goodyear	85338	5773	133	Yes	Yes	Yes	0	In-class	...	21	30	45.6	4.5	43.1
60	Desert Edge High School	Goodyear	85338	2302	81	Yes	Yes	Yes	39	Hybrid	...	25	38	22.2	12.7	57.2
61	Estrella Foothills High School	Goodyear	85338	5773	133	Yes	Yes	Yes	12	In-class	...	21	30	45.6	4.5	43.1
66	Cesar Chavez High School	Laveen	85339	478	18	Yes	Yes	Yes	26	Hybrid	...	24	24	4.1	14.8	74.7
114	Desert Mountain High School	Scottsdale	85259	35	1009	Yes	Yes	Yes	26	Hybrid	...	49	47	74.80	2.90	9.50
121	Imagine Prep Surprise	Surprise	85379	265	14874	No	Yes	Yes	0	Hybrid	...	17	27	53.90	3.90	32.20
124	Valley Vista High School	Surprise	85374	27	608	Yes	No	Yes	104	Hybrid	...	25	32	35.30	9.20	46.70
136	Sierra Linda High School	Tolleson	85353	62	1898	Yes	Yes	Yes	52	Hybrid	...	10	14	3.30	7.90	85.10
163	Maryvale High School	Phoenix	85033	273	14	Yes	Yes	Yes	24	In-class	...	24	24	2.6	4	91
165	Moon Valley High School	Phoenix	85029	3474	107	Yes	Yes	Yes	36	In-class	...	18	30	30.7	8.5	50.7
167	North Canyon High School	Phoenix	85024	1907	64	Yes	Yes	Yes	50	Hybrid	...	33	40	22.7	7.6	61.9
168	North High School	Phoenix	85014	573	24	Yes	Yes	Yes	27	In-class	...	24	24	3.8	6.8	85.2
173	Paradise Valley High	Phoenix	85032	2267	79	Yes	Yes	Yes	21	Hybrid	...	33	40	39.6	3.5	43.6

	School Name	City	Zip Code	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Offers Sports?	Crime Related Data	Teaching and Educational Method	...	Math Score	English Score	Racial% White	Racial% Black	Racial% Hispanic
	School															
176	Pinnacle High School	Phoenix	85050	696	28	Yes	Yes	Yes	18	In-class	...	33	40	74.9	1.8	13.6
177	Roadrunner School	Phoenix	85028	17197	312	No	No	No	8	In-class	...	33	40	48	5.4	31.2
178	Shadow Mountain High School	Phoenix	85028	3930	111	Yes	Yes	Yes	50	Hybrid	...	33	40	53	4.8	33.1
190	Vista Peak	Phoenix	85027	22127	462	No	No	No	2	In-class	...	43	46	21.9	12.5	59.4

27 rows × 22 columns

In [112…
```
df_cleaned['Student Teacher Ratio']
```

Out[112]:
```
0                    25:1
2                    17:1
12                   19:1
15                   23:1
20     Permanently Closed
46                   10:1
48                   22:1
50                   24:1
51                   19:1
56                   21:1
57                   25:1
60                   21:1
61                   25:1
66                   20:1
114                23.1:1
121                N/A1:1
124                26.8:1
136                24.3:1
163                  21:1
165                  20:1
167                  23:1
168                  18:1
173                  18:1
176                  24:1
177                   4:1
178                  15:1
190                   4:1
Name: Student Teacher Ratio, dtype: object
```

In [113…
```
# # Convert categorical attributes into dummy variables
# categorical_columns = ['School Name', 'City', 'Zip Code', 'Offers Sports?']
# df_transformed = pd.get_dummies(df_cleaned, columns=categorical_columns, drop_first=True)

# # Convert boolean columns to integer (0s and 1s)
# boolean_columns = [col for col in df_transformed.columns if df_transformed[col].dtype == 'bool']
# df_transformed[boolean_columns] = df_transformed[boolean_columns].astype(int)

# df_transformed
```

Out[113]:

	National Rank	Arizona Rank	AP Classes	Dual Enrollments	Crime Related Data	Teaching and Educational Method	Mental Health Services	2022N/A2023 Student Enrollment	Student Teacher Ratio	Math Score	...	Zip Code_85308	Zip Code_85323	Zip Code_85326
0	565	23	Yes	Yes	21	Hybrid	Yes	2456	25:1	43	...	0	0	0
2	2423	86	Yes	Yes	48	Hybrid	No	1608	17:1	25	...	0	1	0
12	3772	109	Yes	Yes	36	In-class	Yes	1718	19:1	21	...	0	0	1
15	3378	106	Yes	Yes	46	In-class	Yes	1987	23:1	21	...	0	0	1
20			Permanently Closed	Permanently Closed		Permanently Closed	Permanently Closed		Permanently Closed		...	0	0	0
46	2141	76	Yes	Yes	38	Hybrid	Yes	1568	10:1	43	...	1	0	0
48			No	No	31	In-class	No	632	22:1	49	...	1	0	0
50	1489	49	Yes	Yes	47	Hybrid	No	2157	24:1	18	...	0	0	0
51	4643	119	Yes	Yes	3	Hybrid	No	1170	19:1	30	...	0	0	0
56	2302	81	Yes	Yes	39	Hybrid	No	1786	21:1	25	...	0	0	0
57	5773	133	Yes	Yes	0	In-class	Yes	1253	25:1	21	...	0	0	0
60	2302	81	Yes	Yes	39	Hybrid	No	1786	21:1	25	...	0	0	0
61	5773	133	Yes	Yes	12	In-class	Yes	1253	25:1	21	...	0	0	0
66	478	18	Yes	Yes	26	Hybrid	No	2685	20:1	24	...	0	0	0
114	35	1009	Yes	Yes	26	Hybrid	No	2244	23.1:1	49	...	0	0	0
121	265	14874	No	Yes	0	Hybrid	Yes	293	N/A1:1	17	...	0	0	0
124	27	608	Yes	No	104	Hybrid	No	2531	26.8:1	25	...	0	0	0
136	62	1898	Yes	Yes	52	Hybrid	No	1896	24.3:1	10	...	0	0	0
163	273	14	Yes	Yes	24	In-class	No	2768	21:1	24	...	0	0	0
165	3474	107	Yes	Yes	36	In-class	No	1451	20:1	18	...	0	0	0
167	1907	64	Yes	Yes	50	Hybrid	Yes	1919	23:1	33	...	0	0	0
168	573	24	Yes	Yes	27	In-class	No	2194	18:1	24	...	0	0	0
173	2267	79	Yes	Yes	21	Hybrid	Yes	1865	18:1	33	...	0	0	0
176	696	28	Yes	Yes	18	In-class	Yes	2558	24:1	33	...	0	0	0
177	17197	312	No	No	8	In-class	No	16	4:1	33	...	0	0	0
178	3930	111	Yes	Yes	50	Hybrid	No	1171	15:1	33	...	0	0	0
190	22127	462	No	No	2	In-class	Yes	24	4:1	43	...	0	0	0

27 rows × 74 columns

In [ ]: