Students Performance

The dataset contains information about 2392 high school students. Let's see that exact information is presented in the dataset:

```
In [3]: sql = '''
SELECT COLUMN_NAME
FROM INFORMATION_SCHEMA.COLUMNS
WHERE TABLE_NAME = 'Students_Performance'
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[3]:

	COLUMN_NAME
0	StudentID
1	Age
2	Gender
3	Ethnicity
4	ParentalEducation
5	StudyTimeWeekly
6	Absences
7	Tutoring
8	ParentalSupport
9	Extracurricular
10	Sports
11	Music
12	Volunteering
13	GPA

We can use this information to explore what and how affects student performance. To start, let's see the average, highest and lowest grades of students presented in the dataset.

```
In [4]: sql = '''
SELECT AVG(GPA)Average_Grade,
       MAX(GPA)Highest_Grade,
       MIN(GPA)Lowest_Grade
FROM Students_Performance
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[4]:

	Average_Grade	Highest_Grade	Lowest_Grade
0	1.906186	4.0	0.0

There is an opinion that girls study better than boys. Let's check that:

```
In [5]:  sql = '''
         SELECT g.Gender,
                 COUNT(*) #,
                 AVG(sp.GPA)Average_Grade
         FROM Students_Performance sp
         JOIN Gender g
         ON sp.Gender = g.GenderID
         GROUP BY g.Gender
         '''
         df = pd.read_sql_query(sql, conn)
         df
```

Out[5]:

	Gender	#	Average_Grade
0	Female	1222	1.894225
1	Male	1170	1.918679

Our data does not support this assumption. The difference is too low.

Let's separate students by grades class and check the number of girls and boys in each grade class and see if there is a big difference in any of them:

```
In [6]:  sql = '''
         WITH gradeclass AS
             (SELECT StudentID,
                     CASE
                         WHEN GPA < 2.0
                             THEN 'F'
                         WHEN 2.0 <= GPA and GPA < 2.5
                             THEN 'D'
                         WHEN 2.5 <= GPA and GPA < 3.0
                             THEN 'C'
                         WHEN 3.0 <= GPA and GPA < 3.5
                             THEN 'B'
                         WHEN GPA >= 3.5
                             THEN 'A'
                     END AS GradeClass
             FROM Students_Performance)

         SELECT DISTINCT gc.GradeClass,
                         g.Gender,
                         COUNT(*) OVER (PARTITION BY gc.GradeClass, g.Gender)
         FROM Students_Performance sp
         JOIN gradeclass gc
         ON sp.StudentID = gc.StudentID
         JOIN Gender g
         ON sp.Gender = g.GenderID
         '''
         df = pd.read_sql_query(sql, conn)
         df
```

Out[6]:

	GradeClass	Gender	
0	A	Female	41
1	A	Male	36
2	B	Female	120
3	B	Male	124
4	C	Female	193
5	C	Male	192
6	D	Female	209
7	D	Male	203
8	F	Female	659
9	F	Male	615

We only see significant difference in the last group.

Let's check if weekly study time influences performance. Compare the average weekly time for each grade class:

In [7]:
```
sql = '''
WITH gradeclass AS
    (SELECT StudentID,
            CASE
                WHEN GPA < 2.0
                    THEN 'F'
                WHEN 2.0 <= GPA and GPA < 2.5
                    THEN 'D'
                WHEN 2.5 <= GPA and GPA < 3.0
                    THEN 'C'
                WHEN 3.0 <= GPA and GPA < 3.5
                    THEN 'B'
                WHEN GPA >= 3.5
                    THEN 'A'
            END AS GradeClass
        FROM Students_Performance)

SELECT gc.GradeClass,
       AVG(sp.StudyTimeWeekly)Average_StudyTimeWeekly
FROM Students_Performance sp
JOIN gradeclass gc
ON sp.StudentID = gc.StudentID
GROUP BY gc.GradeClass
ORDER BY gc.GradeClass
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[7]:

	GradeClass	Average_StudyTimeWeekly
0	A	12.891884
1	B	11.357646
2	C	10.101910
3	D	9.899795
4	F	9.138707

The assumption was supported. Students with A-grades study more. Let's check if parental education and parental support affect students' performances.

In [8]:
```
sql = '''
SELECT pe.ParentalEducation,
       AVG(sp.GPA)Average_Grade
FROM Students_Performance sp
JOIN ParentalEducation pe
ON sp.ParentalEducation = pe.ParentalEducationID
GROUP BY pe.ParentalEducation, sp.ParentalEducation
ORDER BY sp.ParentalEducation
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[8]:

	ParentalEducation	Average_Grade
0	None	1.893045
1	High School	1.944022
2	Some College	1.929881
3	Bachelor	1.809082
4	Higher	1.815812

In [9]:
```
sql = '''
SELECT ps.ParentalSupport,
       AVG(sp.GPA)Average_Grade
FROM Students_Performance sp
JOIN ParentalSupport ps
ON sp.ParentalSupport = ps.ParentalSupportID
GROUP BY ps.ParentalSupport, sp.ParentalSupport
ORDER BY sp.ParentalSupport
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[9]:

	ParentalSupport	Average_Grade
0	None	1.540128
1	Low	1.755700
2	Moderate	1.884246
3	High	2.042409
4	Very High	2.191545

We can make conclusion that parental education does not really affect children's performance. Students, whose parents have bachelor or higher degree, even have lower grades, but the difference is too low to make some serious conclusions.
But we can conclude that parental support really means a lot for kids and the higher support - the higher the grade.

Let's now compare grades of students who participate or do not participate in activities like sport, music and volunteering.

In [10]:
```
sql = '''
SELECT Sports,
       AVG(GPA)Average_Grade
FROM Students_Performance
GROUP BY Sports
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[10]:

	Sports	Average_Grade
0	1	1.986381
1	0	1.871240

In [11]:
```
sql = '''
SELECT Music,
       AVG(GPA)Average_Grade
FROM Students_Performance
GROUP BY Music
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[11]:

	Music	Average_Grade
0	1	2.041664
1	0	1.872969

In [12]:
```
sql = '''
SELECT Volunteering,
       AVG(GPA)Average_Grade
FROM Students_Performance
GROUP BY Volunteering
'''
df = pd.read_sql_query(sql, conn)
df
```

Out[12]:

	Volunteering	Average_Grade
0	1	1.913089
1	0	1.904899

We see that students who participate in any of these activities have higher grades than those who do not.
Let's check the average grade of a student who does not participate in these or any other extracurricular activities:

```
In [13]: sql = '''
         SELECT  COUNT(GPA),
                 AVG(GPA)Average_Grade
         FROM Students_Performance
         WHERE Sports = 0
               AND Music = 0
               AND Volunteering = 0
               AND Extracurricular = 0
         '''
         df = pd.read_sql_query(sql, conn)
         df
```

Out[13]:

		Average_Grade
0	679	1.752377

We see that there are 679 student in the dataset who do not participate in any extracurricular activities and their grades are lower than the average.

```
In [ ]:
```