Результати виконання ТЗ

1.Завантаження та підготовка даних

Встановлення всіх необхідних бібліотек

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import f_oneway
import scipy.stats as stats
import seaborn as sns
```

Імпортовано всі необхідні бібліотеки та модулі

Зчитуємо дані з csv файлу та створюємо DataFrame для подальшої роботи

In [5]:
```
data=pd.read_csv('Student_performance_data _.csv')
print(data.columns)
```

```
Index(['StudentID', 'Age', 'Gender', 'Ethnicity', 'ParentalEducation',
       'StudyTimeWeekly', 'Absences', 'Tutoring', 'ParentalSupport',
       'Extracurricular', 'Sports', 'Music', 'Volunteering', 'GPA',
       'GradeClass'],
      dtype='object')
```

Виводимо назви стовпців

Перевіряємо чи існують відсутні значення

In [6]:
```
message = "Відсутніх значень немає. Рухаємось до наступного кроку" if not data.isnull().any().any() else "Існують відсут
print(message)
```

```
Відсутніх значень немає. Рухаємось до наступного кроку
```

Перевірка на наявність порожніх значень. Порожніх значень в наборі даних немає

1.Завантаження та підготовка даних

Перевірка на наявність аномалій відбувається методом міжквартильного інтервалу. Ті значення, які знаходяться поза межами 1 та 3 квартиля, на відстані більше ніж півтора міжквартильного інтервалу - рахуються аномаліями та будуть видалені з датасету

In [8]:
```python
# Функція для виявлення аномалій і видалення їх з DataFrame
def remove_anomalies(df):
    Q1 = df['Age'].quantile(0.25)
    Q3 = df['Age'].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Видалення аномалій
    df = df[(df['Age'] >= lower_bound) & (df['Age'] <= upper_bound)]

    return df

# Видалення аномалій з DataFrame
cleaned_data = remove_anomalies(data)

print("Очищений DataFrame без аномалій:")
print(cleaned_data)
```

```
Очищений DataFrame без аномалій:
```

За результатами очищення, було виявлено що від самого початку, датасет не містив аномальних значень по данним полям. Розмір набору даних не змінився

За таким принципом, відбувається очищення по полям: Age, StudyTimeWeekly, Absences, GPA.

2. Описова статистика

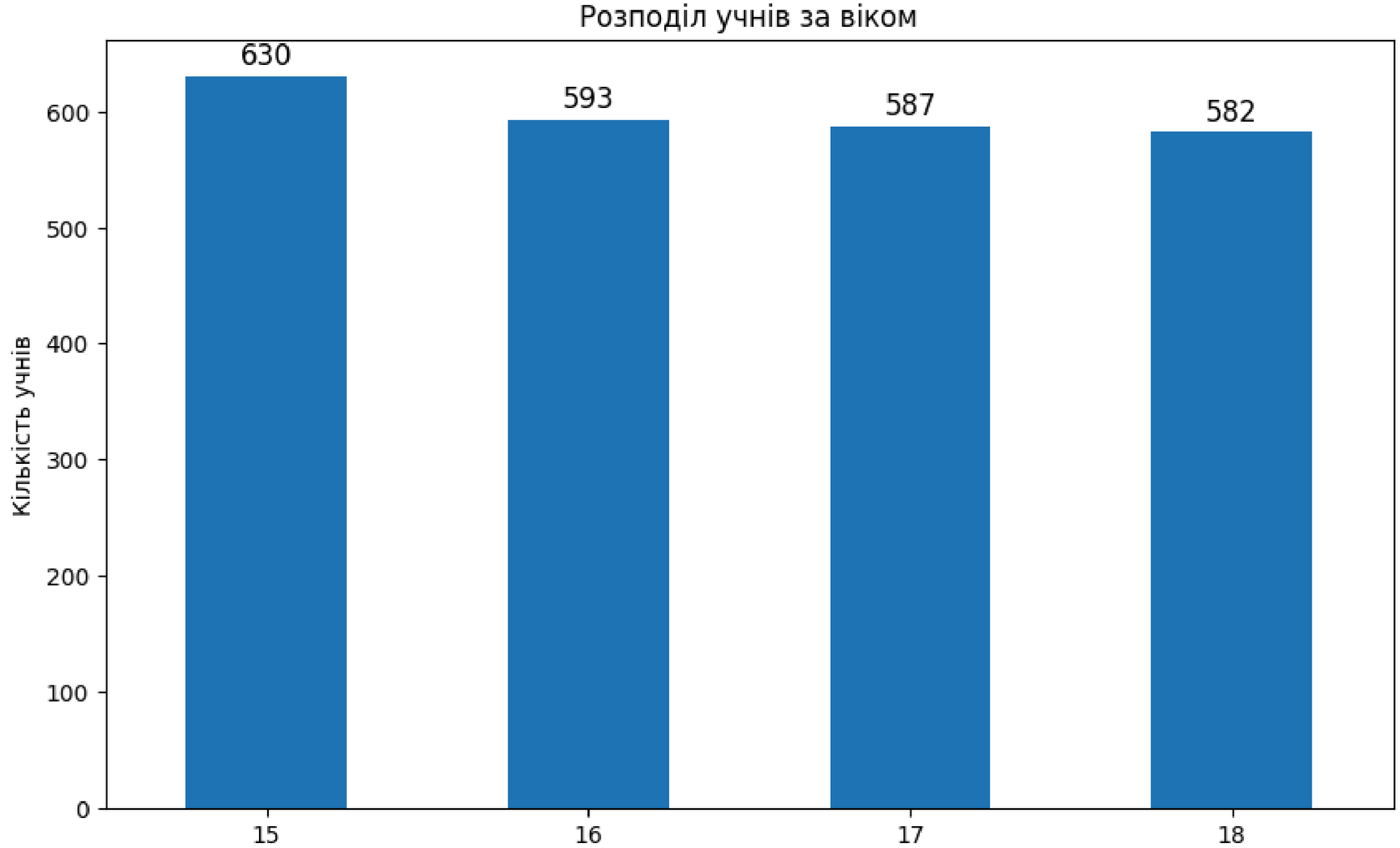
In [14]:
```
print(cleaned_data[['Age','StudyTimeWeekly','Absences','GPA','GradeClass']].describe())
```

	Age	StudyTimeWeekly	Absences	GPA	GradeClass
count	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000
mean	16.468645	9.771992	14.541388	1.906186	2.983696
std	1.123798	5.652774	8.467417	0.915156	1.233908
min	15.000000	0.001057	0.000000	0.000000	0.000000
25%	15.000000	5.043079	7.000000	1.174803	2.000000
50%	16.000000	9.705363	15.000000	1.893393	4.000000
75%	17.000000	14.408410	22.000000	2.622216	4.000000
max	18.000000	19.978094	29.000000	4.000000	4.000000

Для решти полів описова статистика не застосовується, оскільки вони є якісними і не підлягають опису

2. Описова статистика. Розподіл учнів за віком, статтю, етнічністю та рівнем освіти батьків

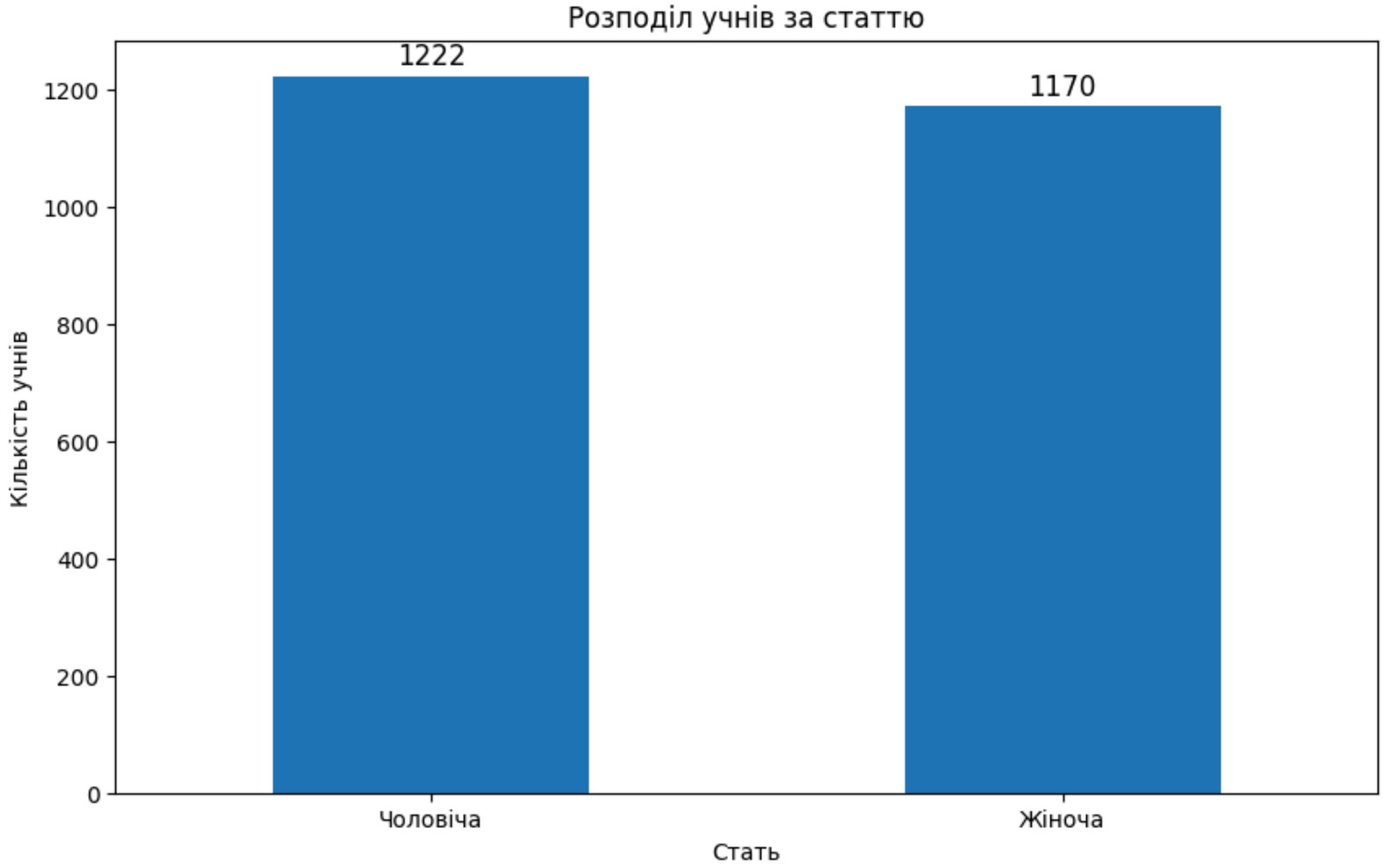
Для описової статистики були використано гістограми



Можна помітити, що найчисельніша група - 15 років. Найменша група - 18 років

2. Описова статистика. Розподіл учнів за віком, статтю, етнічністю та рівнем освіти батьків

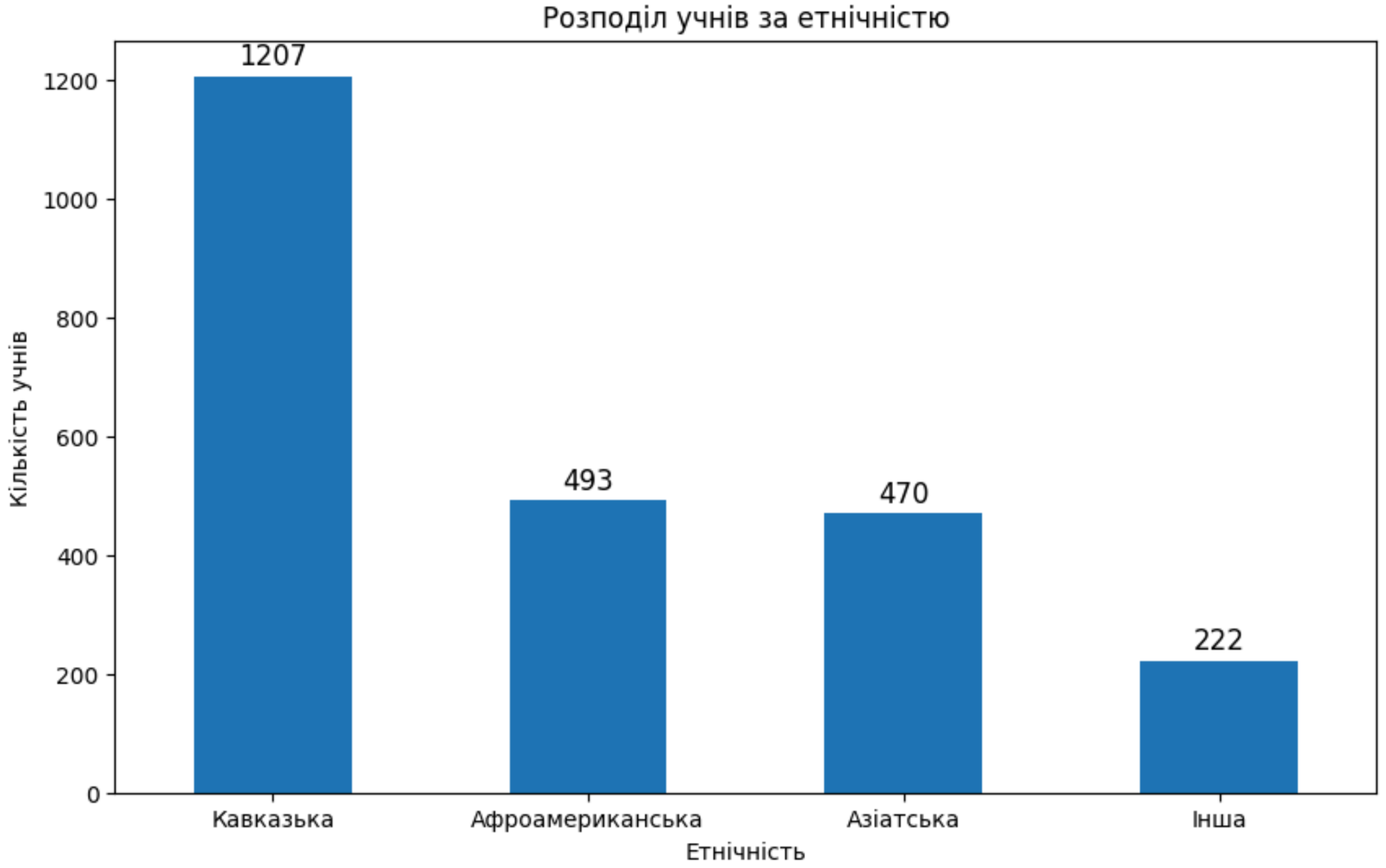
Для описової статистики були використано гістограми



Можна помітити, що у вибірці чоловіків більше ніж жінок. Чоловіків 1222 а жінок 1170.

2. Описова статистика. Розподіл учнів за віком, статтю, етнічністю та рівнем освіти батьків

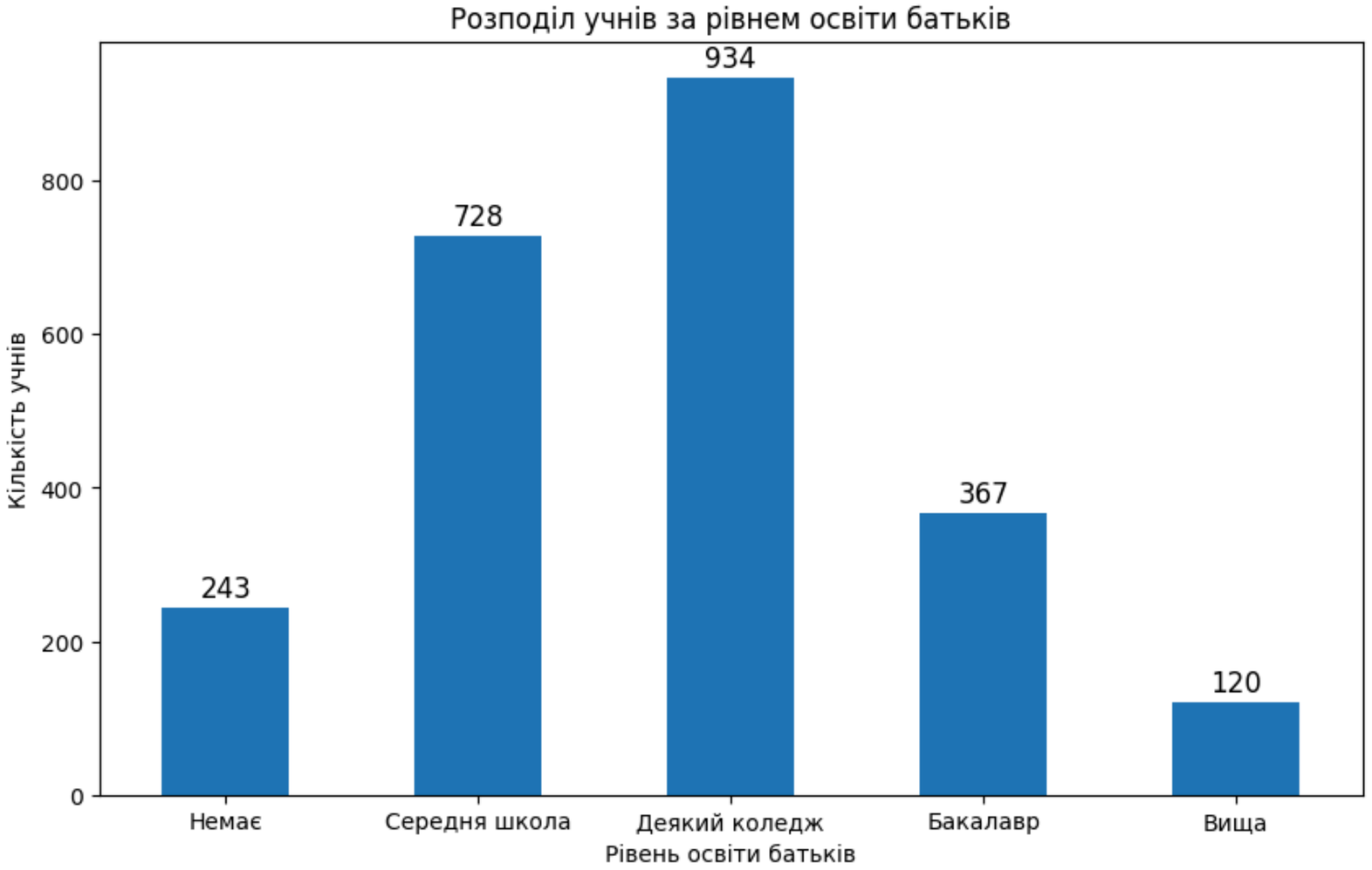
Для описової статистики були використано гістограми



Розподіл за етнічністтю наступний.
Найчисельніша група у вибірці - кавказька.
Найменш чисельна - Інша

2. Описова статистика. Розподіл учнів за віком, статтю, етнічністю та рівнем освіти батьків

Для описової статистики були використано гістограми



Розподіл учнів за рівнем освіти наступний. Можна побачити, що найбільше батьків мають освіту - Деякий коледж. Найменше - Вища.

3. Аналіз навчальних звичок. Кореляція між навчальним часом на тиждень та кількістю пропусків

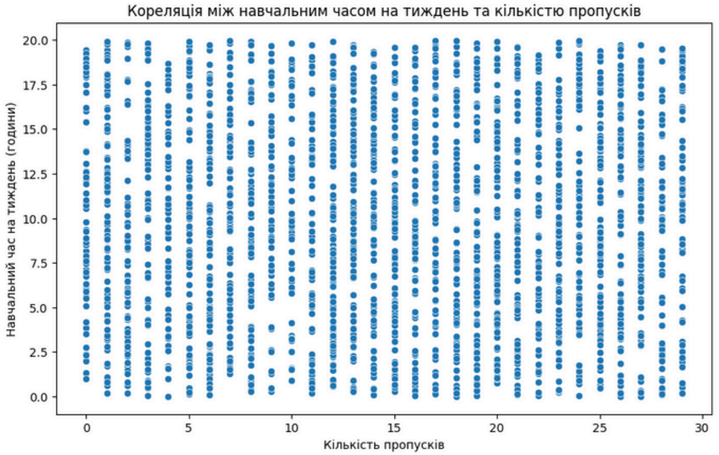
Вивчити кореляцію між навчальним часом на тиждень та кількістю пропусків

```
corr = stats.pearsonr(cleaned_data['StudyTimeWeekly'],cleaned_data['Absences']).statistic
print(f"Кореляція між навчальним часом на тиждень та кількістю пропусків: {round(corr,4)}, тобто відсутня")
```

```
кореляція між навчальним часом на тиждень та кількістю пропусків: 0.0093, тобто відсутня
```

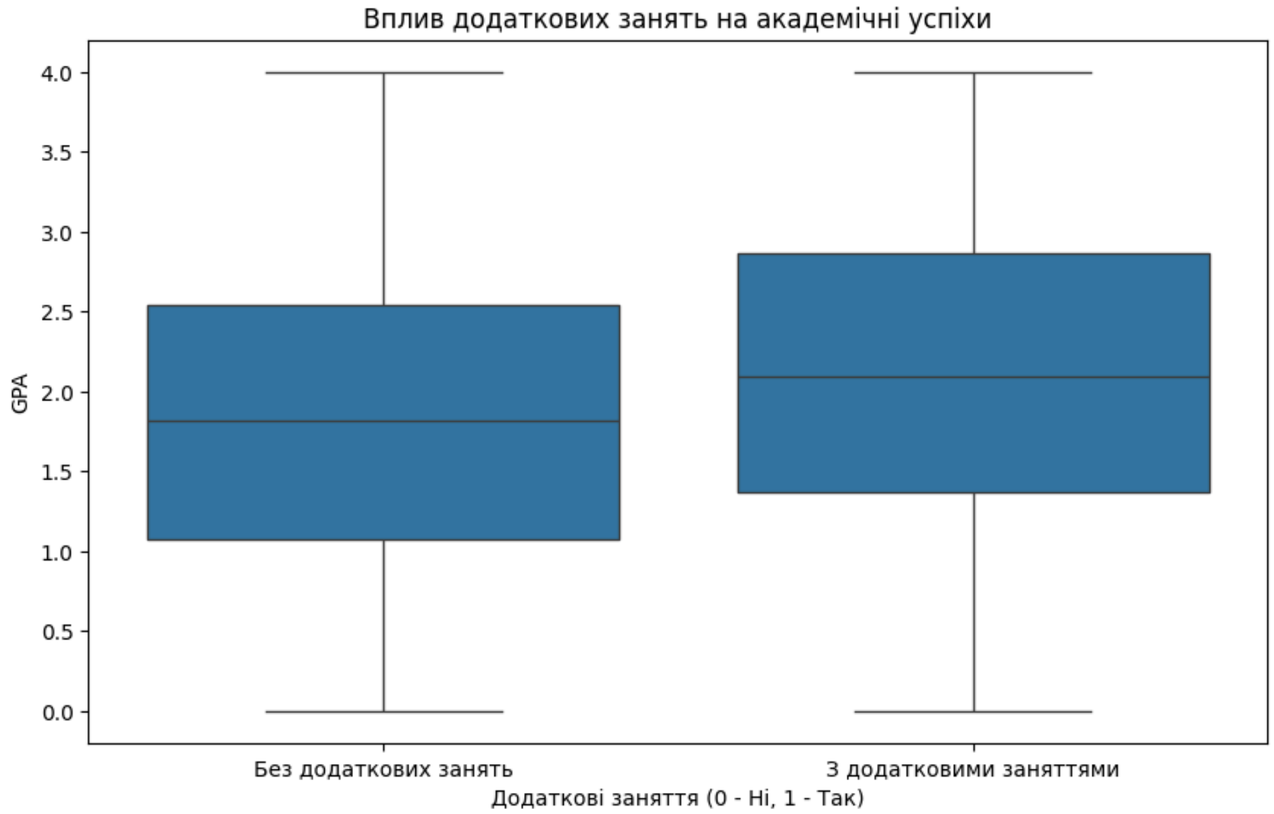
Також графічно можна побачити що зв'язок відсутній

```
import seaborn as sns
import matplotlib.pyplot as plt

# Побудова графіка кореляції
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Absences', y='StudyTimeWeekly', data=cleaned_data)
plt.title('Кореляція між навчальним часом на тиждень та кількістю пропусків')
plt.xlabel('Кількість пропусків')
plt.ylabel('Навчальний час на тиждень (години)')
plt.show()
```



Коефіцієнт кореляції становить 0,0093 - тобто кореляція відсутня

3. Аналіз навчальних звичок. Аналіз впливу додаткових занять на академічні успіхи



Середній GPA для учнів з додатковими заняттями: 2.11

Середній GPA для учнів без додаткових занять: 1.82

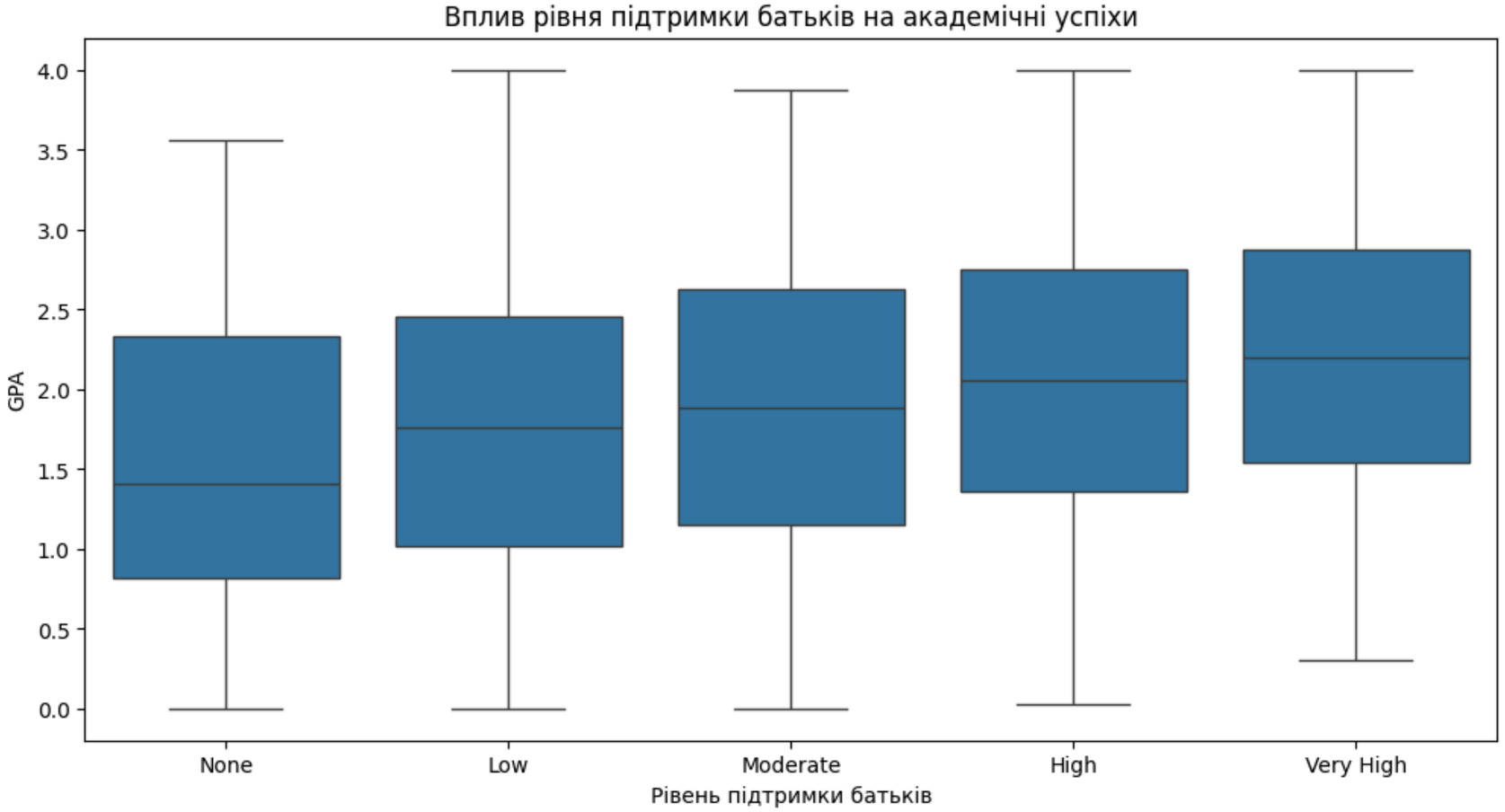
Гіпотеза H0 про відсутність різниці в середніх значеннях GPA між учнями з додатковими заняттями та без них відхиляється. Результати показують, що додаткові заняття мають статистично значущий вплив на академічні успіхи учнів.

```
t-статистика: 7.17
p-значення: 0.0000
Різниця в GPA між учнями з додатковими заняттями та без них є статистично значущою.
```

4. Аналіз участі батьків. Як рівень підтримки батьків впливає на академічні результати учнів

	ParentalSupport	GPA
0	0	1.540128
1	1	1.755700
2	2	1.884246
3	3	2.042409
4	4	2.191545

Спочатку було розраховано середні академічні результати учнів в залежності від рівня підтримки батьків де 0 - низький, а 4 - високий



Візуалізація результатів

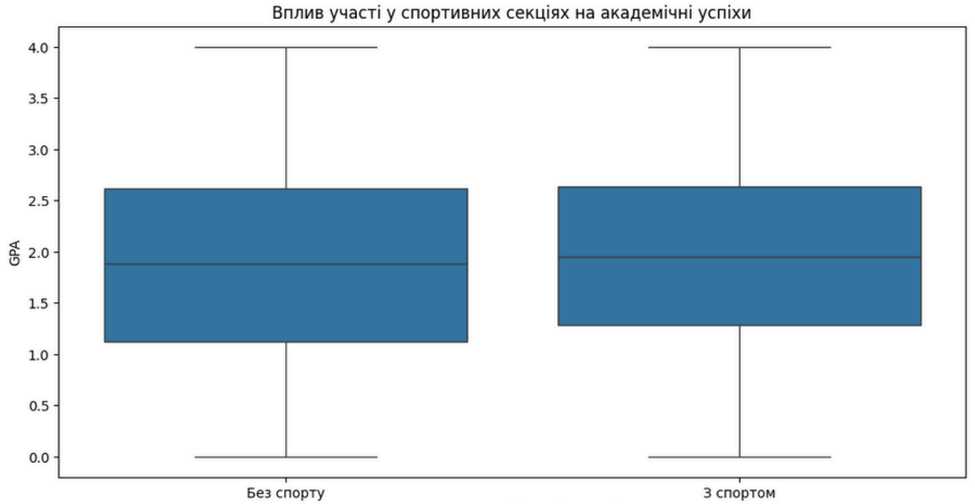
4. Аналіз участі батьків. Як рівень підтримки батьків впливає на академічні результати учнів

```
F-статистика: 22.72
p-значення: 0.0000
Різниця в GPA між групами з різним рівнем підтримки батьків є статистично значущою.
```

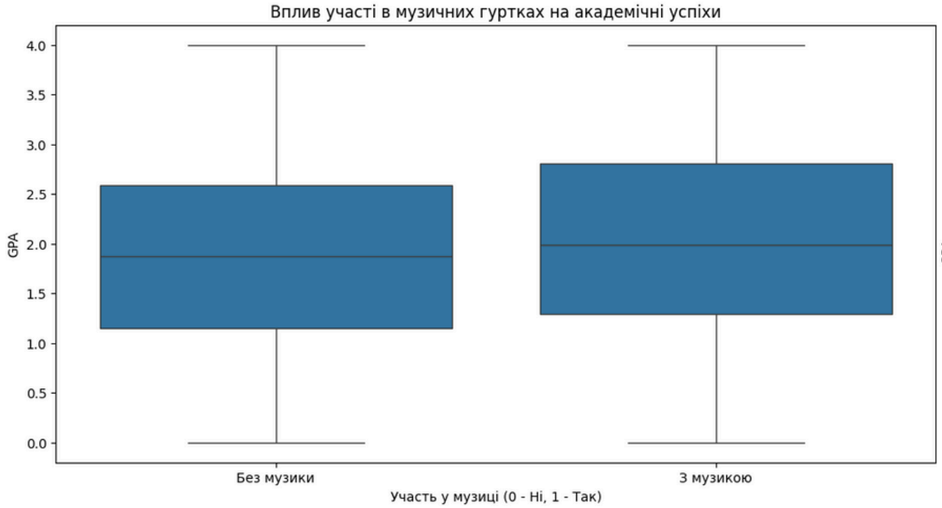
Гіпотеза H0 про відсутність різниці в середніх значеннях GPA між групами з різним рівнем підтримки батьків відхиляється. Результати показують, що рівень підтримки батьків має статистично значущий вплив на академічні успіхи учнів

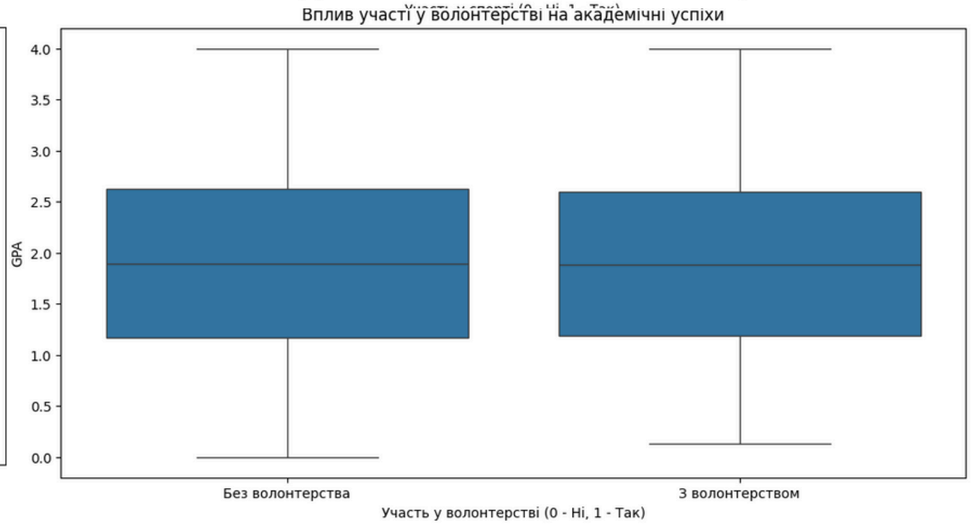
5. Аналіз позашкільної діяльності. Вплив участі у позашкільних заходах (спортивні секції, музичні гуртки, волонтерство) на GPA

```
Середній GPA для учнів, які займаються спортом: 1.99
Середній GPA для учнів, які не займаються спортом: 1.87

Середній GPA для учнів, які займаються музикою: 2.04
Середній GPA для учнів, які не займаються музикою: 1.87

Середній GPA для учнів, які займаються волонтерством: 1.91
Середній GPA для учнів, які не займаються волонтерством: 1.90
```



Графіічно можна побачити волонтерство не впливає на GPA. Проведемо t-тест





5. Аналіз позашкільної діяльності. Вплив участі у позашкільних заходах (спортивні секції, музичні гуртки, волонтерство) на GPA

```
t-статистика для спортивних секцій: 2.83
p-значення для спортивних секцій: 0.0046

t-статистика для музичних гуртків: 3.59
p-значення для музичних гуртків: 0.0003

t-статистика для волонтерства: 0.16
p-значення для волонтерства: 0.8735

Різниця в GPA між учнями, які займаються спортом, і тими, хто не займається спортом, є статистично значущою.

Різниця в GPA між учнями, які займаються музикою, і тими, хто не займається музикою, є статистично значущою.

Різниця в GPA між учнями, які займаються волонтерством, і тими, хто не займається волонтерством, не є статистично значущою.
```

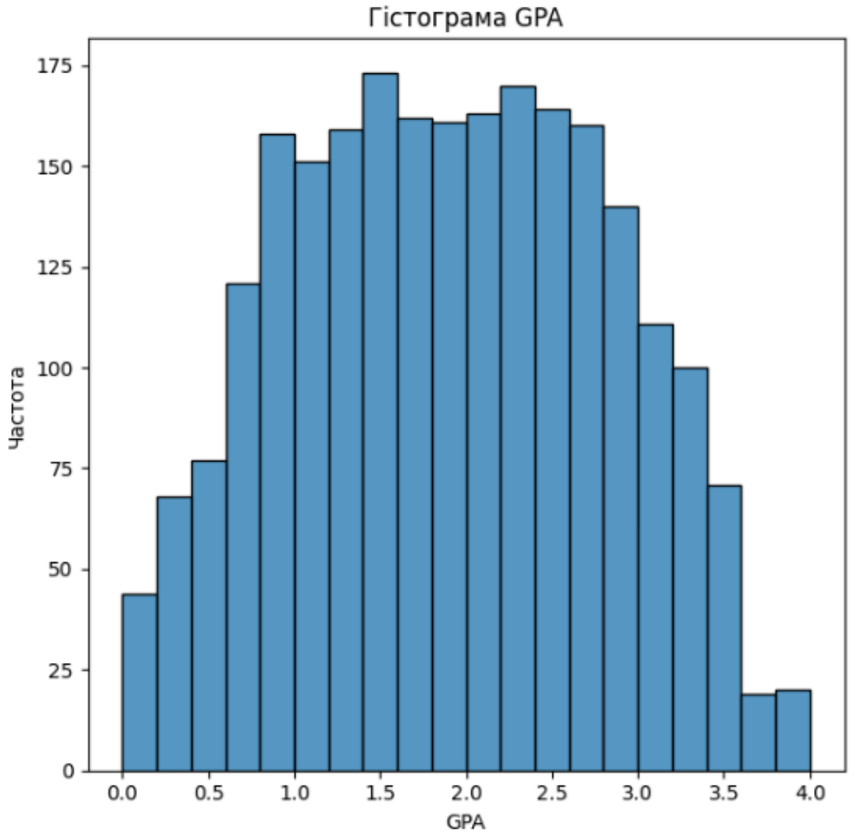
Спортивні секції: Оскільки p-значення (0.0046) менше 0.05, гіпотеза H0 відхиляється. Різниця в GPA є статистично значущою.

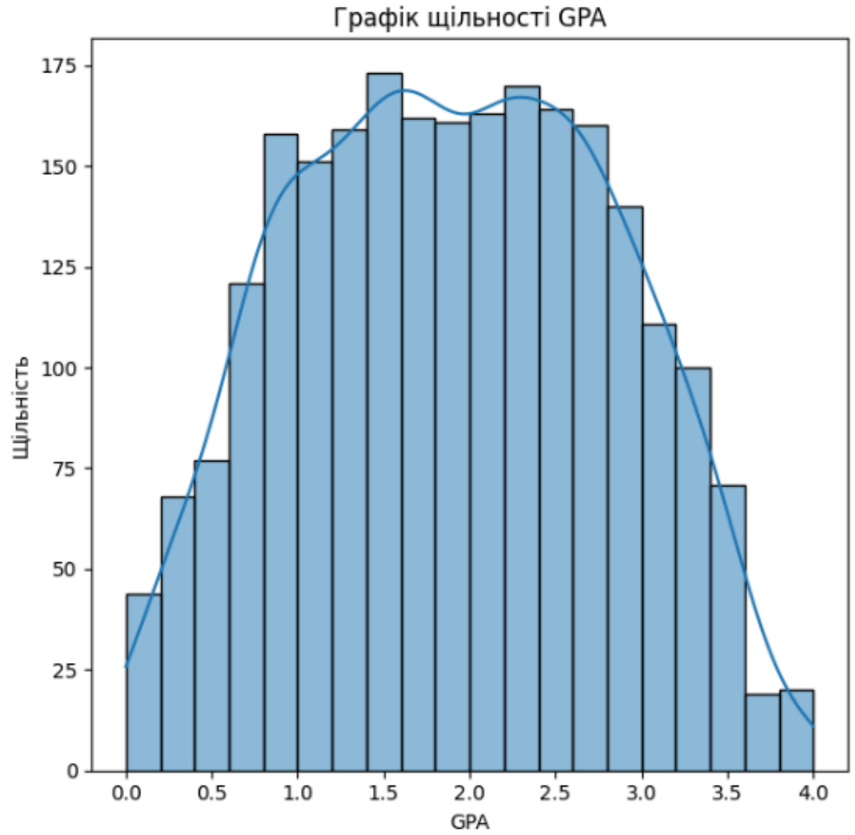
Музичні гуртки: Оскільки p-значення (0.0003) менше 0.05, гіпотеза H0 відхиляється. Різниця в GPA є статистично значущою.

Волонтерство: Оскільки p-значення (0.8735) більше 0.05, гіпотеза H0 не відхиляється. Різниця в GPA не є статистично значущою. Тобто волонтерство ніяк не впливає на GPA.

6.Аналіз академічних результатів. Розподіл GPA серед учнів

```
Основні статистичні показники GPA:
count    2392.000000
mean        1.906186
std         0.915156
min         0.000000
25%         1.174803
50%         1.893393
75%         2.622216
max         4.000000
```

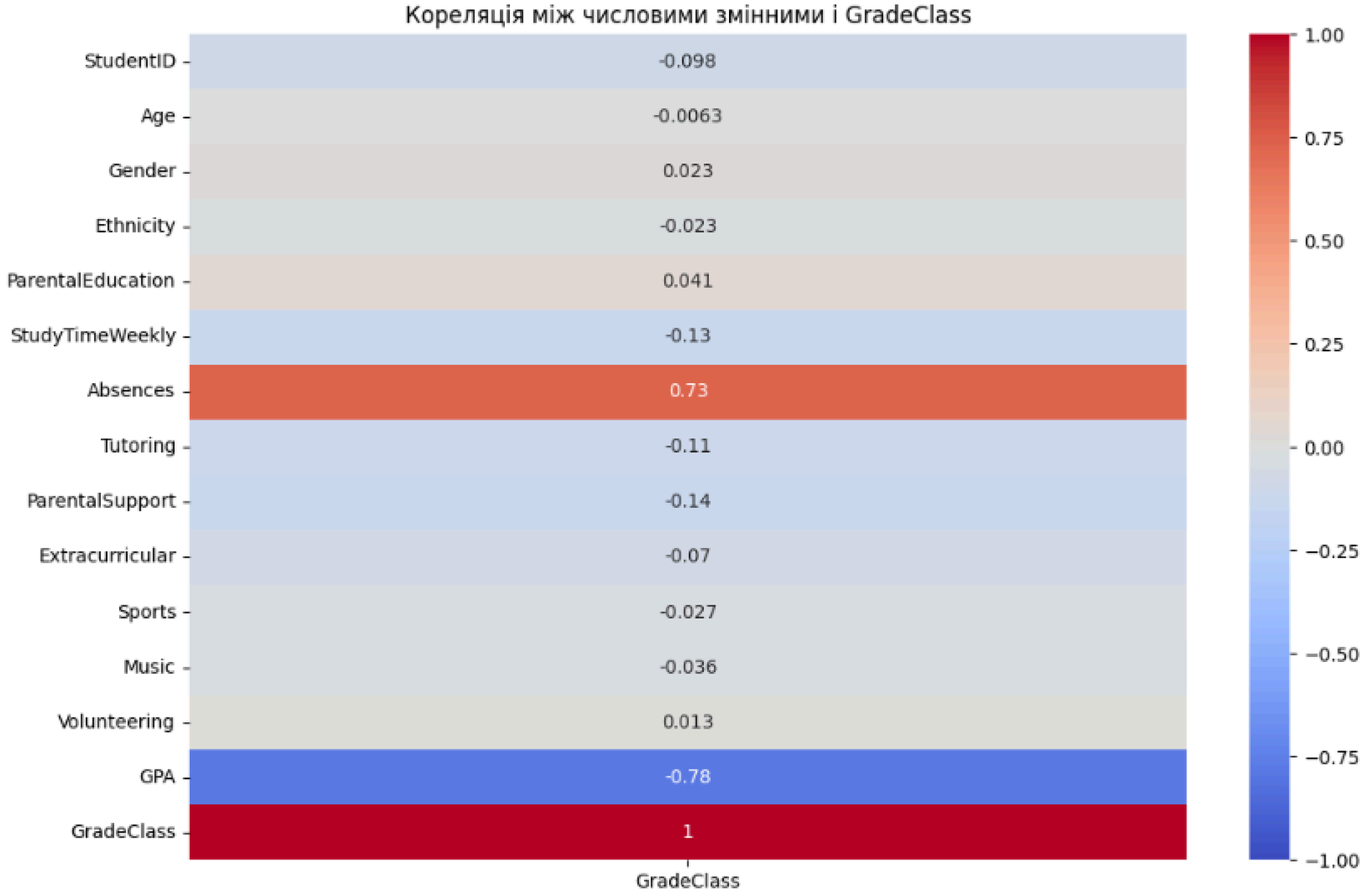





```
Статистика тесту Шапіро-Уїлка: 0.9838
p-значення тесту Шапіро-Уїлка: 0.0000
Розподіл GPA не є нормально розподіленим.
```

Для перевірки нормальності розподілу GPA використовувався тест Шапіро-Уілка. Результати показали що ряд не є нормально розподіленим

6.Аналіз академічних результатів. основні фактори, що впливають на класифікацію оцінок (GradeClass)



Для кращого відображення, було побудовано графік для відображення кореляцій

Тільки Absences та GPA мають високі показники парної кореляції. перевіримо значущість цих показників

6.Аналіз академічних результатів. основні фактори, що впливають на класифікацію оцінок (GradeClass)

Значущі коефіцієнти кореляції (p < 0.05) для вибраних пар змінних:

	Variable 1	Variable 2	Correlation	t-statistic	p-value
0	GradeClass	GPA	-0.782835	-61.506268	0.0
1	Absences	GradeClass	0.728633	52.008880	0.0

Гіпотеза H0 про не значущість парних коефіцієнтів кореляції між GradeClass - GPA та GradeClass - Absences відхиляється. Результати показують, дані коефіцієнти кореляції є статистично значущими