

Разведочный анализ данных

Выполнил: Бабак З. О.

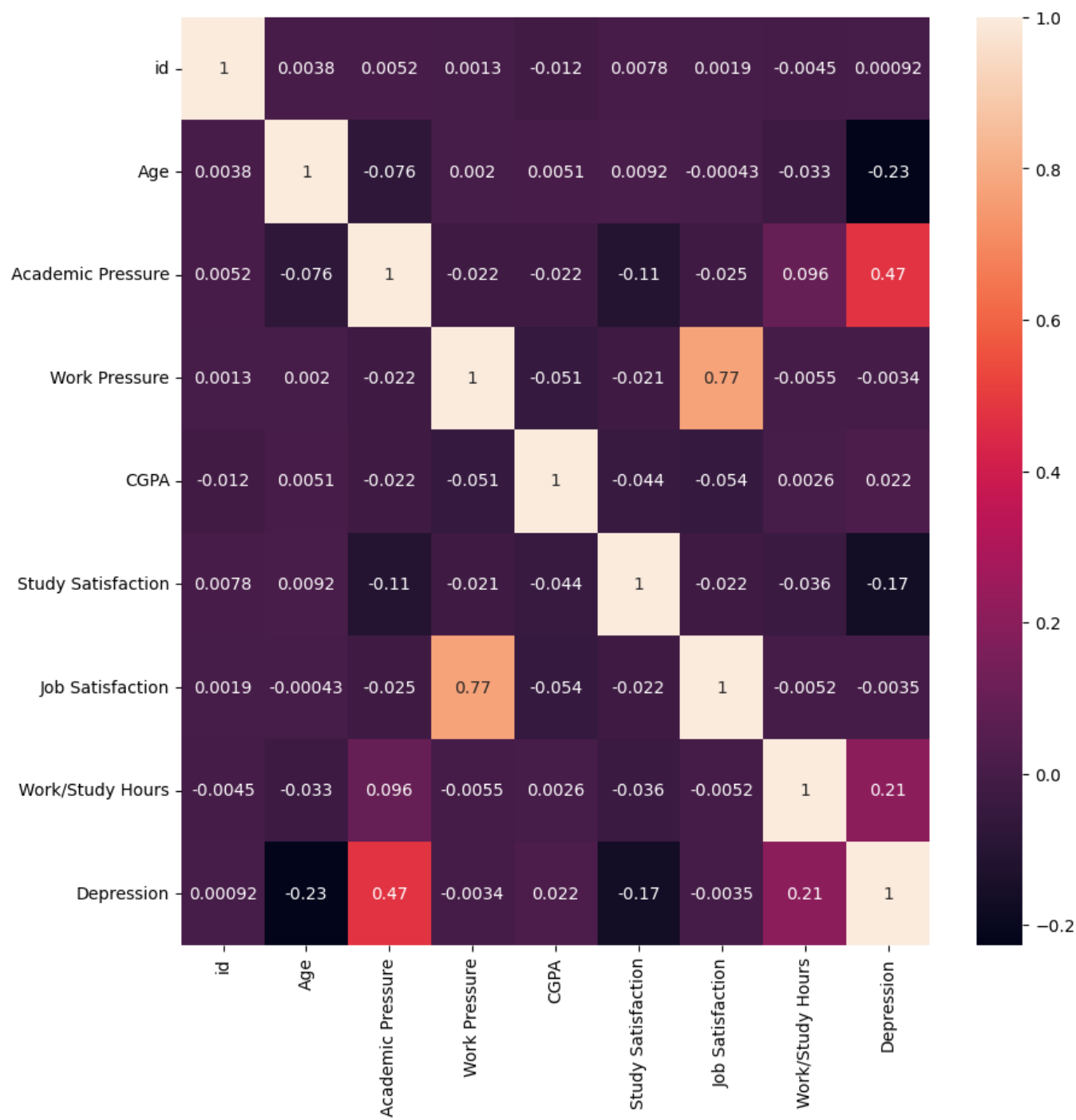
Учебная группа: Б23-215

Основной анализ

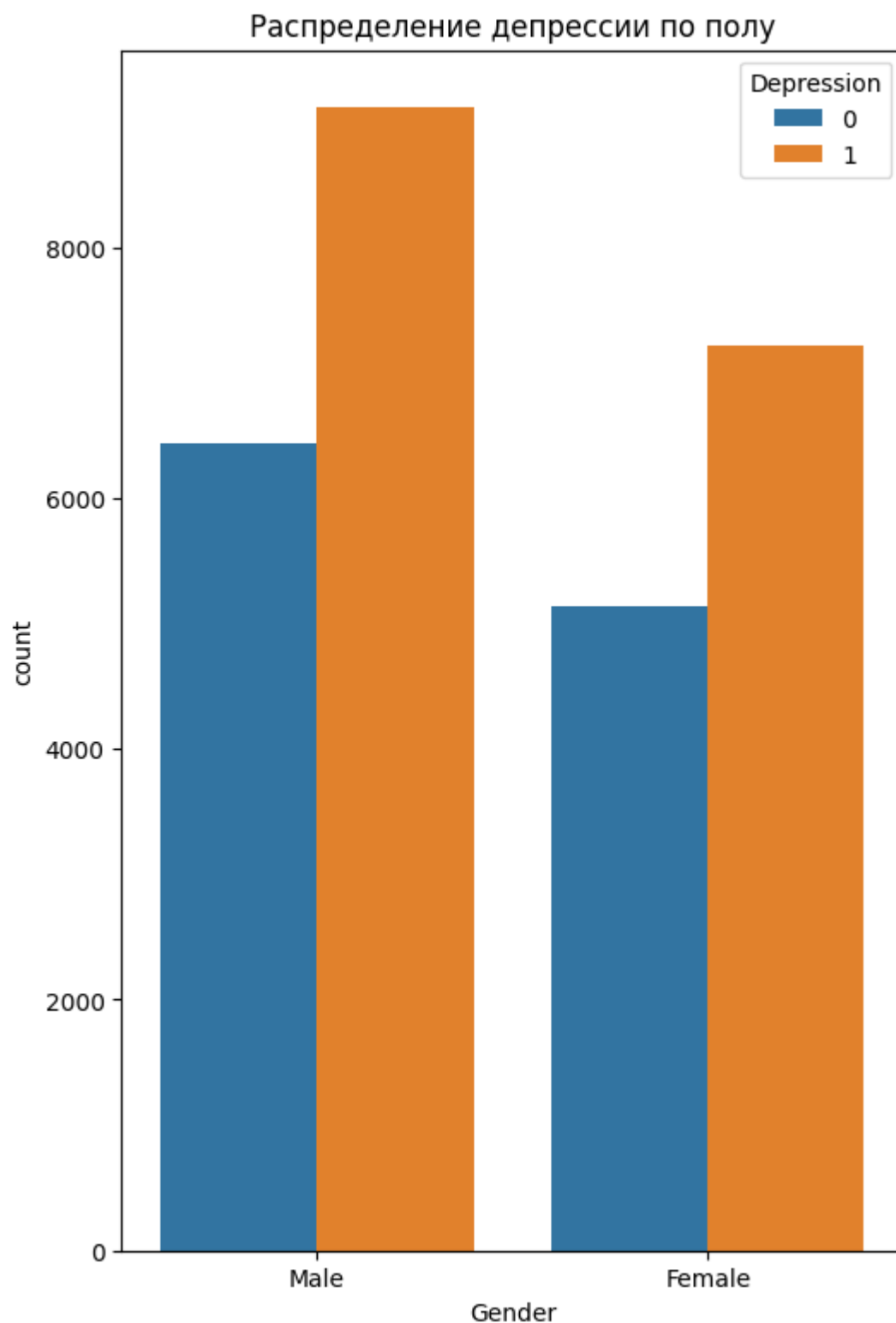
Посмотрим, сколько людей из общего числа имеют депрессию:

Число человек с депрессией: 16336, это примерно 58.55 процентов от общего числа.

Матрица корреляции



Видим, что наибольшая корреляция у целевой переменной "Depression" с параметром академическое давление. Среди остальных параметров, наблюдается некоторая корреляция с отношением часов учебы к часам работы, а также обратная корреляция с возрастом и удовлетворенностью учебой.



По данному графику видим, что число мужчин с депрессией выше. Но стоит проверить количественное соотношение мужчин и женщин в наборе данных.

Число мужчин: 15547

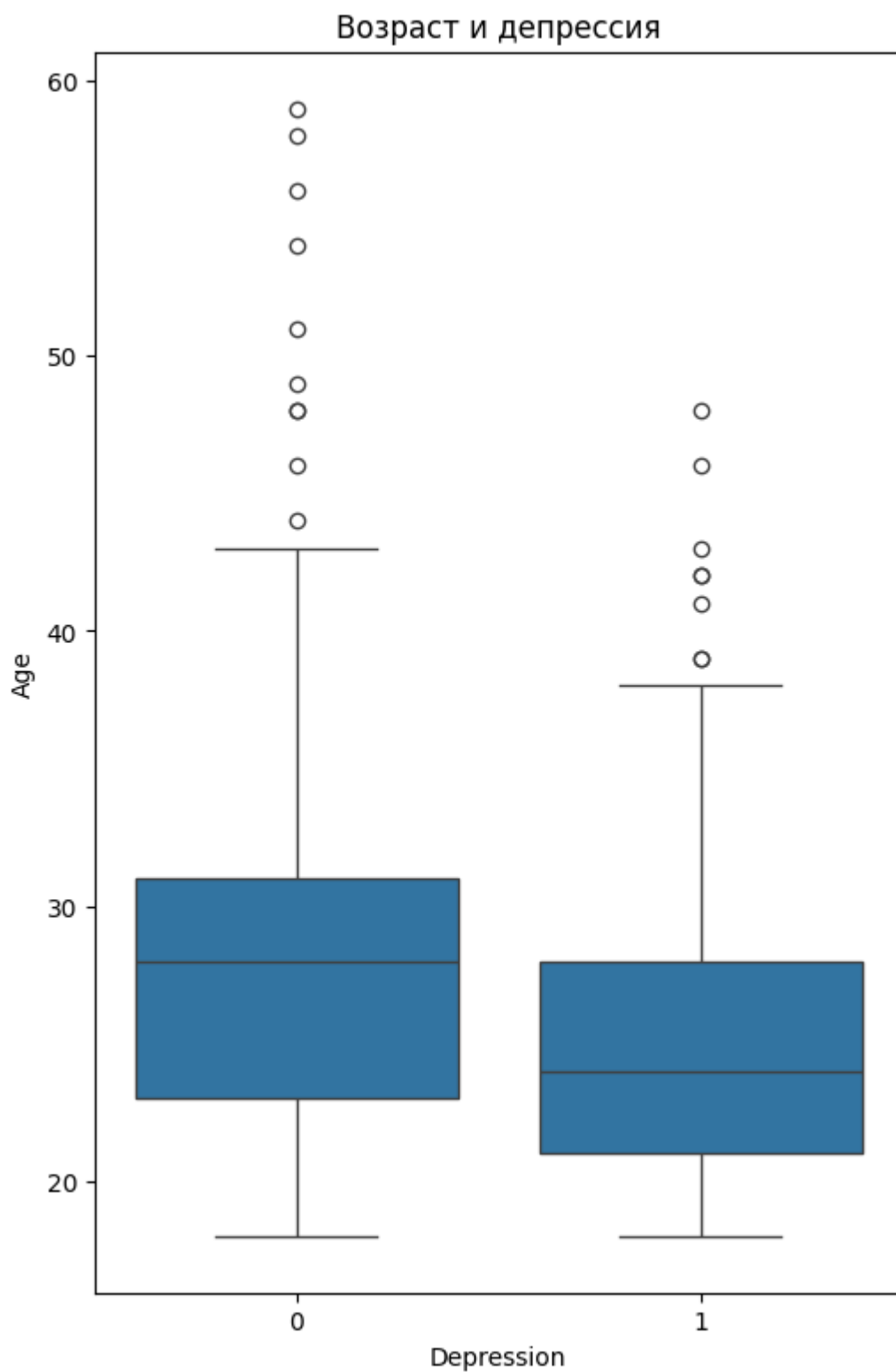
Число женщин: 12354

Женщин примерно на 20% меньше, поэтому посчитаем отношение людей с депрессией к общему числу для каждого пола для более объективного анализа.

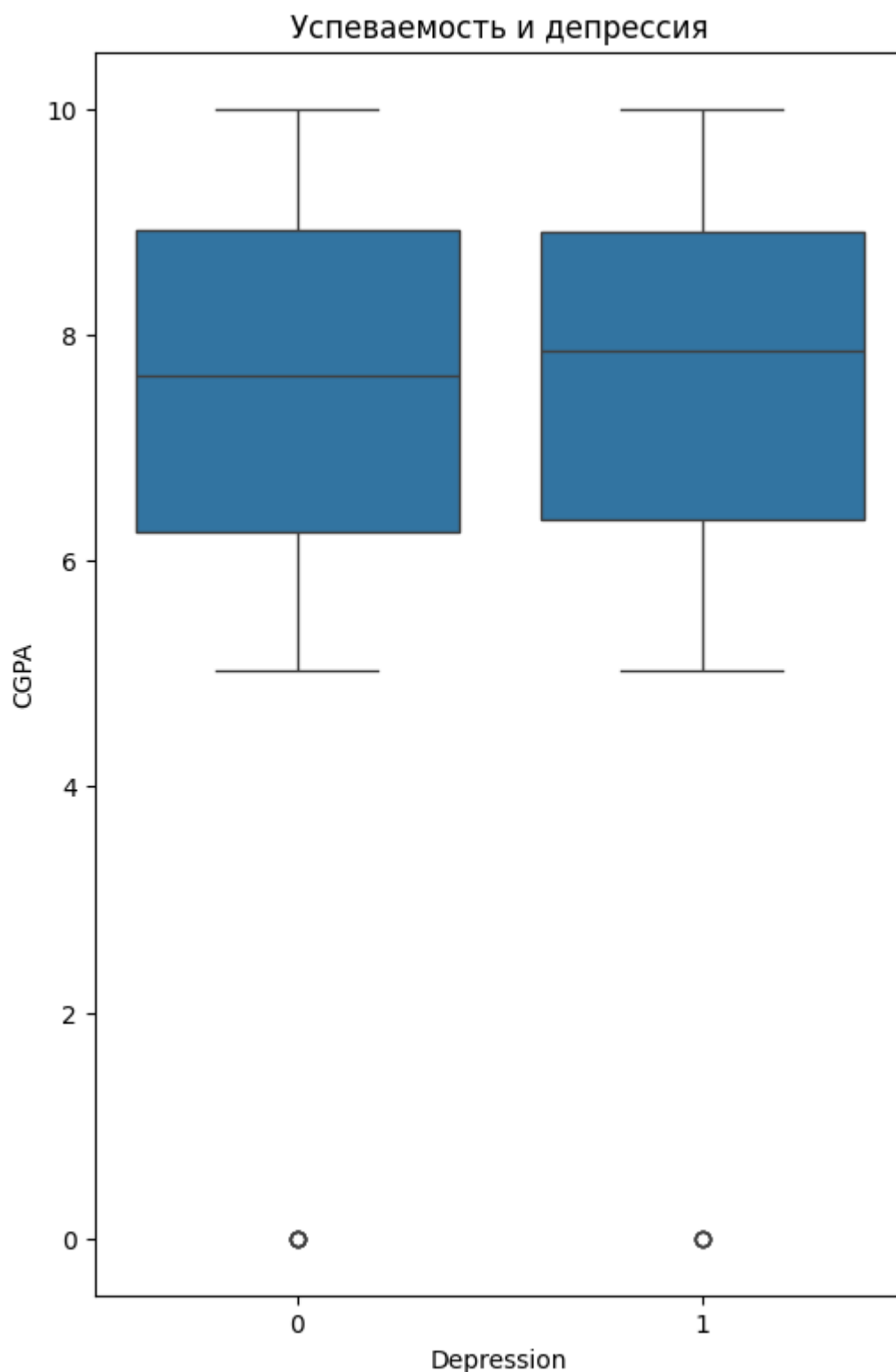
Доля депрессии среди мужчин: 0.5862867434231684

Доля депрессии среди женщин: 0.5845070422535211

Видим, что в процентном соотношении количество мужчин и женщин с депрессией примерно одинаковое.



По данному ящику с усами заключаем, что медиана возраста для людей с депрессией около 24-25 лет, для людей без депрессии — около 28 лет. Возраста выше 38 и 43 лет являются выбросами для людей с депрессией и без депрессии соответственно.



Можем сделать вывод, что для людей с депрессией и без депрессии такие статистические параметры успеваемости студентов, как медиана, верхний и нижний квантили примерно одинаковы, а также не имеется выбросов.

Попробуем посмотреть, как влияют академическая нагрузка и нагрузка на работе на наличие депрессии. Посчитаем и сравним среднее значение каждого параметра отдельно для людей с

депрессией и без.

Среднее давление на учебе у людей с депрессией: 3.693131733594515

Среднее давление на учебе у людей без депрессии: 2.361608300907912

Среднее давление на работе у людей с депрессией: 0.0003060724779627816

Среднее давление на работе у людей без депрессии: 0.0006052745352356247

Таким образом, уровень нагрузки на работе фактически не влияет на наличие депрессии. А вот учебная нагрузка у людей с депрессией примерно в полтора раза выше, чем у людей без депрессии.

Выясним, сколько представителей различных профессий имеется в наборе данных. Profession

'Civil Engineer' 1

'Content Writer' 2

'Digital Marketer' 3

'Educational Consultant' 1

'UX/UI Designer' 1

Architect 8

Chef 2

Doctor 2

Entrepreneur 1

Lawyer 1

Manager 1

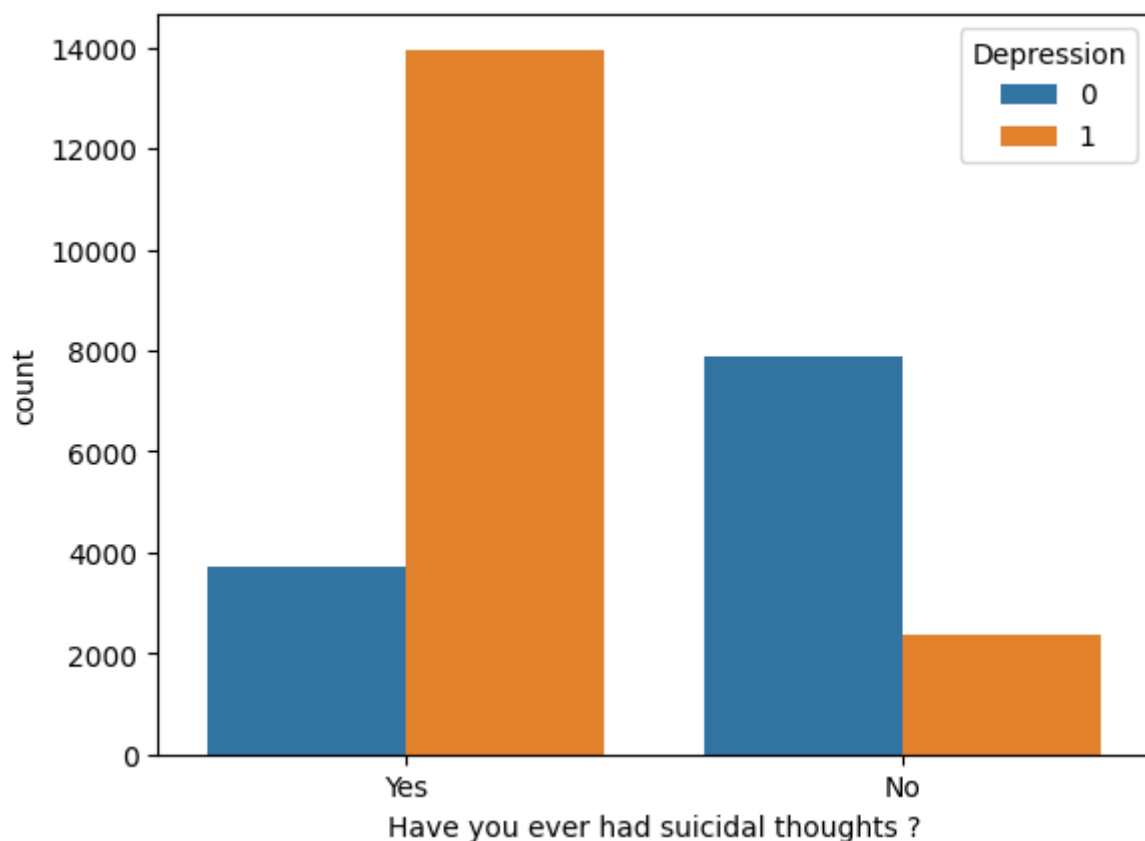
Pharmacist 2

Student 27870

Teacher 6

Name: Profession, dtype: int64

Как видим, датасет практически не содержит профессий помимо студента, следовательно изучать зависимости целевой переменной от профессии нецелесообразно.



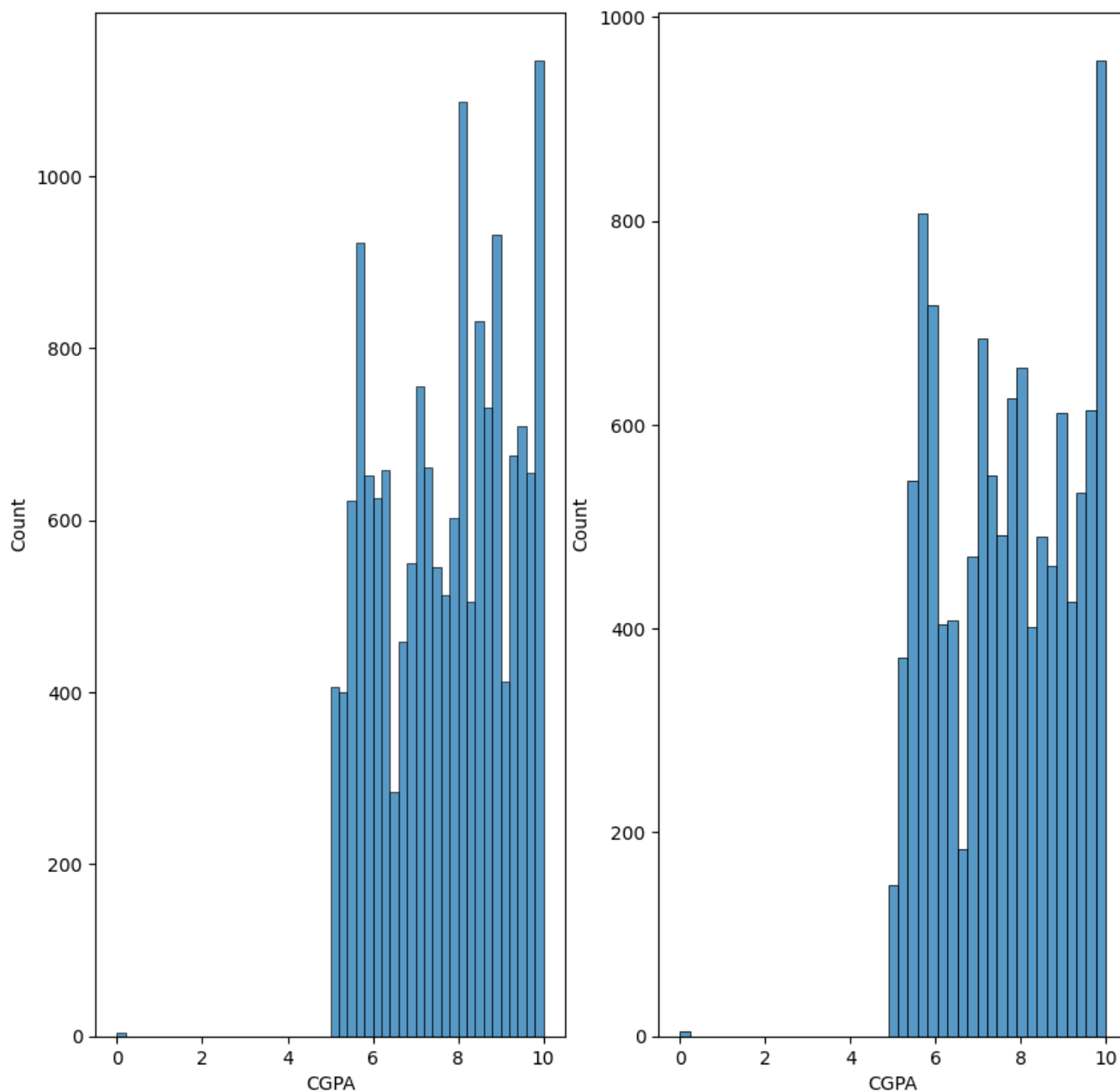
Таким образом, у примерно 14000 человек из чуть больше чем 16000 с депрессией возникали суицидальные мысли.

Статистический анализ

Проверим несколько статистических гипотез.

Гипотеза 1

Посмотрим, как отличаются средние баллы у студентов с депрессией и без. Выдвинем предположение: успеваемость студентов с депрессией выше, чем у студентов без депрессии. Для использования t-критерия нужно, чтобы данные имели нормальное распределение. Посмотрим на распределение на графике, а также реализуем проверку в коде (аналогично для последующих гипотез).



Как видно из графика, значения имеют распределение, отличное от нормального. Поэтому для проверки гипотезы будем использовать U-тест Манна-Уитни.

Тест Шапиро-Уилка на нормальность:

Группа с депрессией: p-value = 0.0000

Группа без депрессии: p-value = 0.0000

Тест Левена на равенство дисперсий: p-value = 0.0001

Использованный тест: U-тест Манна-Уитни

p-value: 0.0001

Результат: Отвергаем нулевую гипотезу

Вывод: Успеваемость студентов с депрессией СТАТИСТИЧЕСКИ ЗНАЧИМО выше

Гипотеза 2

Выясним, как влияет соотношение часов работы и учебы на наличие депрессии. Выдвинем предположение: коэффициент Work/Study hours у людей с депрессией выше.

Тест Шапиро-Уилка на нормальность:

Группа с депрессией: p-value = 0.0000

Группа без депрессии: p-value = 0.0000

Тест Левена на равенство дисперсий: p-value = 0.0000

Использованный тест: U-тест Манна-Уитни

p-value: 0.0000

Результат: Отвергаем нулевую гипотезу

Вывод: соотношение часов работы и учебы у студентов с депрессией СТАТИСТИЧЕСКИ ЗНАЧИМО выше

Гипотеза 3

Выясним, как влияет удовлетворение учебой на наличие депрессии. Выдвинем предположение: коэффициент Study satisfaction у людей с депрессией выше.

Тест Шапиро-Уилка на нормальность:

Группа с депрессией: p-value = 0.0000

Группа без депрессии: p-value = 0.0000

Тест Левена на равенство дисперсий: p-value = 0.0003

Использованный тест: U-тест Манна-Уитни

p-value: 1.0000

Результат: Принимаем нулевую гипотезу

Вывод: Нет статистически значимых различий в удовлетворенности учебой

Гипотеза 4

Выясним, у какого пола после 30 лет депрессия встречается чаще. Выдвинем предположение: после 30 лет у мужчин депрессия встречается чаще.

Кросс-таблица (пол vs депрессия):

Depression 0 1

Gender

Female 1511 1093

Male 2094 1401

Использованный тест: Хи-квадрат

p-value: 0.1451

Результат не значим. Нет оснований отвергать нулевую гипотезу.

Как видим, тест показал, что результат не значим, но мужчин с депрессией после 30 больше, чем женщин. Поверим соотношение долей.

Доля мужчин после 30 лет с депрессией: 0.4008583690987124

Доля женщин после 30 с депрессией: 0.4197388632872504

Доли практически совпадают, что и привело к тому, что тест не выявил статистической значимости.