Отчёт по анализу алкогольной зависимости у студентов

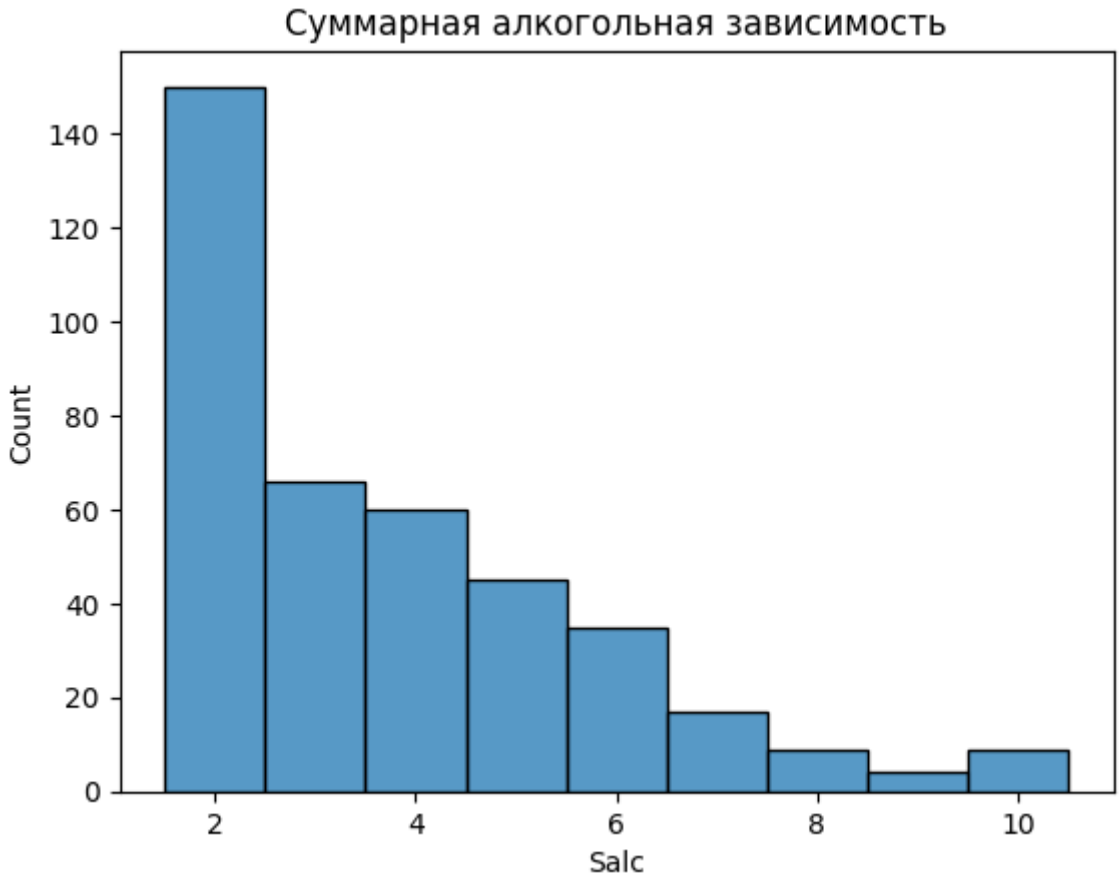
Описание данных

Датасет взят отсюда. Исследуемый датасет содержит информацию о 395 студентах и включает следующие ключевые переменные:

- **Демографические данные**: пол, возраст, адрес, размер семьи, статус родителей
- **Семейные факторы**: образование родителей, профессии, отношения в семье
- **Академические показатели**: оценки (G1, G2, G3), время учебы, пропуски
- **Потребление алкоголя**: будни (Dalc) и выходные (Walc)
- **Дополнительные переменные**: хобби, интернет, отношения и др.

Созданы производные переменные:

- `G` = G1 + G2 + G3 (суммарная оценка)
- `Salc` = Dalc + Walc (суммарное потребление алкоголя в неделю)



Предварительный анализ данных

Распределение ключевых переменных

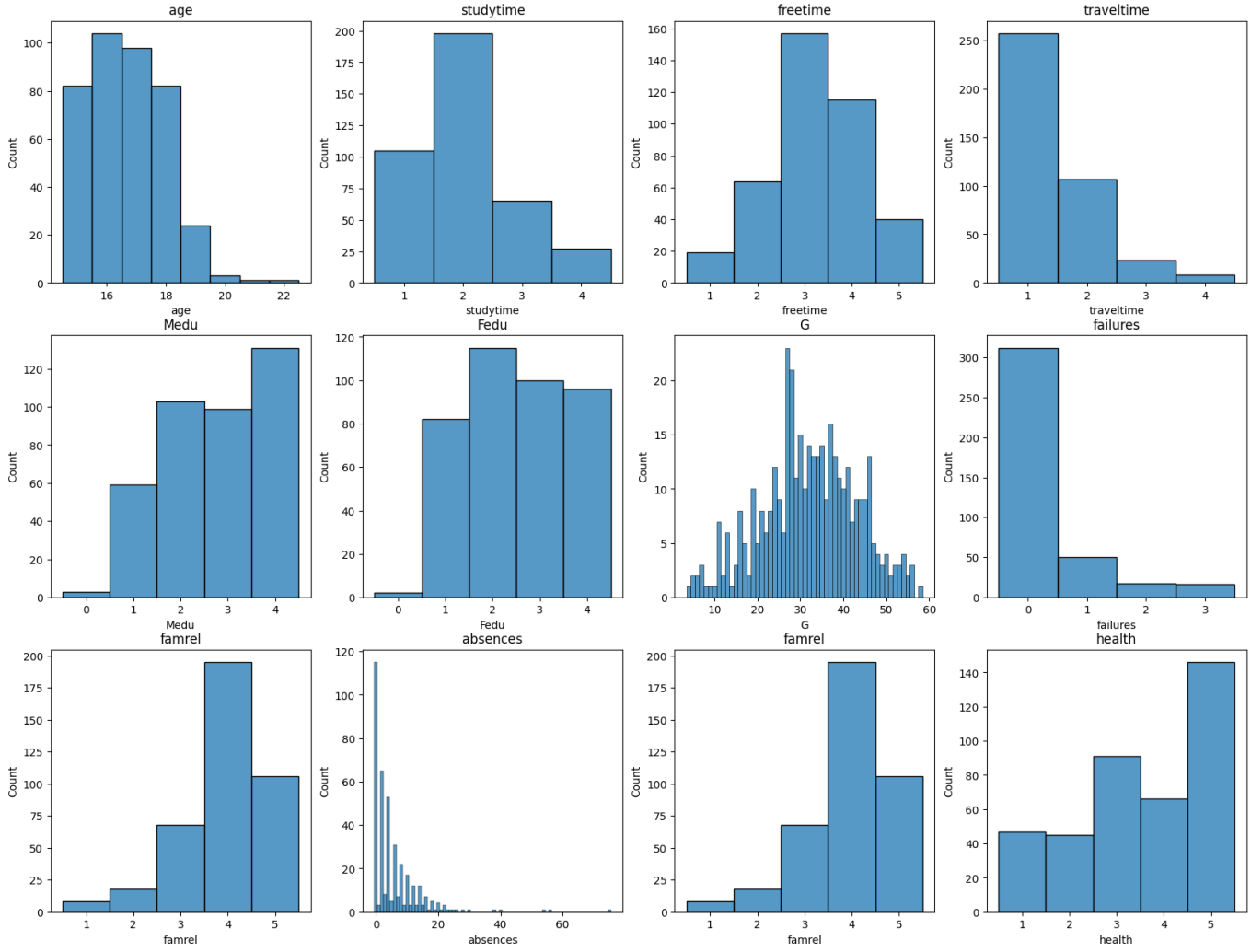
Демография:

- Пол: 208 женщин (52.7%), 187 мужчин (47.3%)
- Размер семьи: GT3 (281), LE3 (114)
- Статус родителей: живут вместе (354), отдельно (41)

Алкогольное потребление:

- Распределение Salc похоже на экспоненциальное
- Потребление выше на выходных по сравнению с буднями

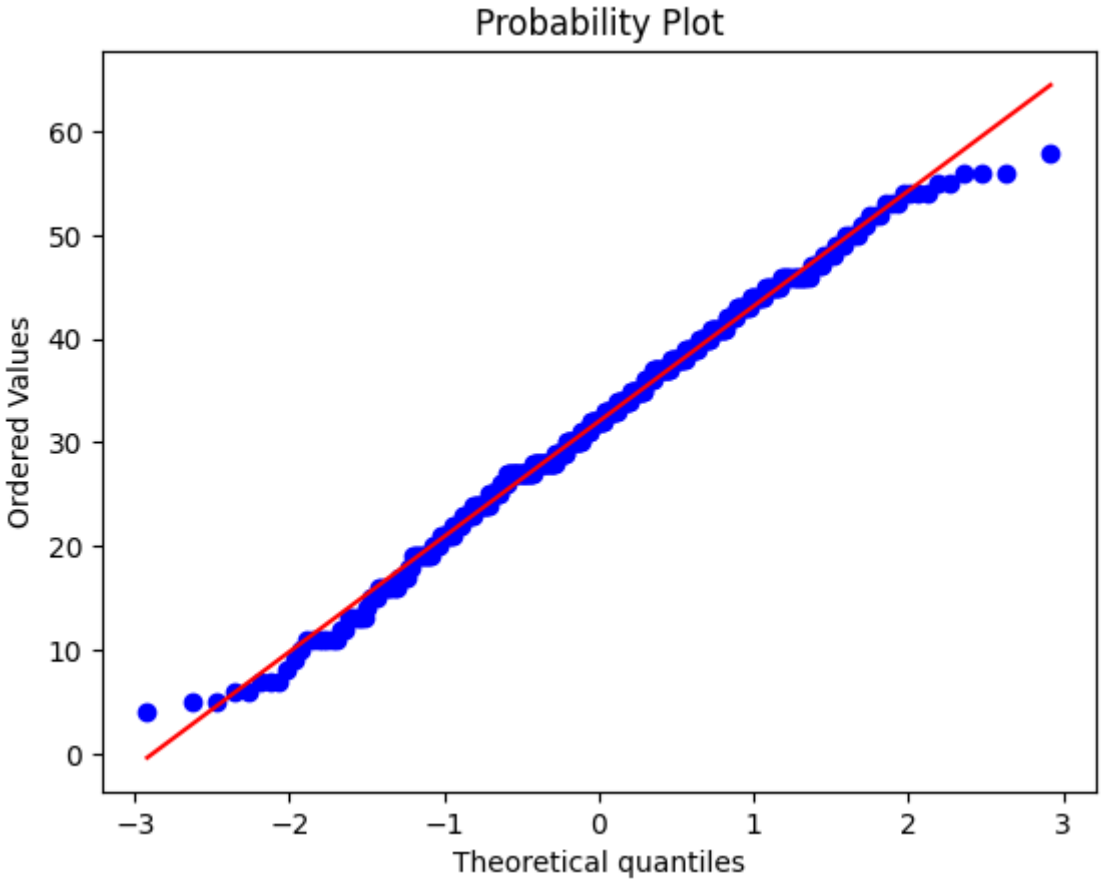
Картичонки:



Проверка нормальности распределения оценок

Метод пристального взгляда: похоже

Хотя левый хвост облегчен, а правый утяжелён:



- **Среднее**: 32.04, **стандартное отклонение**: 11.09

- Правило трёх сигм выполняется:
    - $±1σ$: 68.4% данных (ожидается 68.3%)
    - $±2σ$: 95.9% данных (ожидается 95.4%)
    - $±3σ$: 100% данных (ожидается 99.7%)

Статистические тесты нормальности:

- Тест Шапиро-Уилка: p-value = 0.0505 > 0.05
- Тест D'Agostino: p-value = 0.1578 > 0.05

Оба теста не отвергают гипотезу о нормальном распределении.

Статистические гипотезы

1. Влияние статуса родителей на потребление алкоголя

Гипотезы:

- $H_0: E(Salc|T) = E(Salc|A)$ (нет различий)
- $H_1: E(Salc|T) ≤ E(Salc|A)$ (дети разведенных родителей пьют больше)

Метод: Тест Манна-Уитни (распределение не нормальное)

Результаты:

- Вместе: среднее = 3.77, ст.откл. = 1.94
- Отдельно: среднее = 3.83, ст.откл. = 2.35
- p-value = 0.7111

Вывод: Нет статистически значимых различий ($p > 0.05$). Принимаем H_0.

2. Влияние алкоголя на успеваемость

Корреляционный анализ:

- Корреляция Salc и G: -0.089676

Сравнение групп:

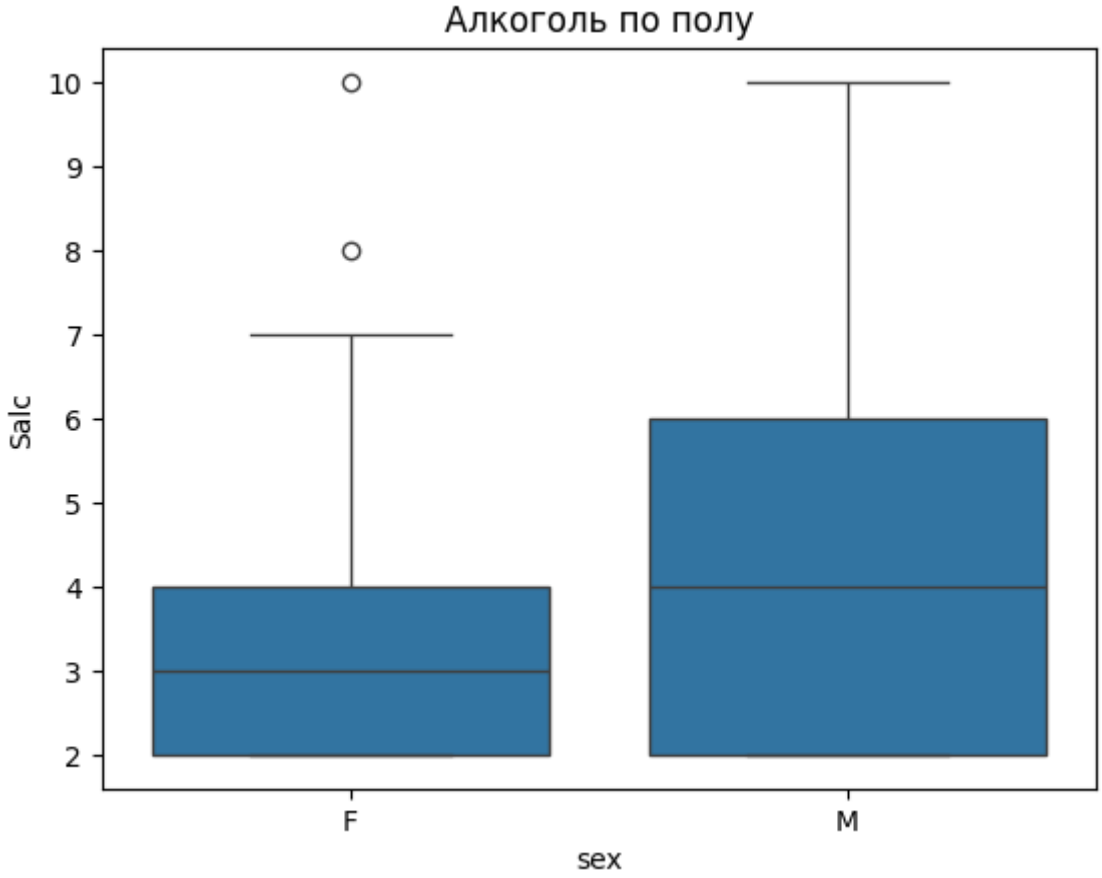
- Высокое потребление vs низкое потребление
- t-test p-value: 0.1286

Вывод: Нет статистически значимых различий в успеваемости.

3. Гендерные различия в потреблении алкоголя

Гипотезы: Мужчины пьют больше женщин

Метод: Пристального взгляда на барплот:



Вывод: Мужчины пьют побольше.

4. Влияние дополнительных занятий на успеваемость

Результаты:

- С доп. занятиями: 32.48
- Без доп. занятий: 31.58
- Разница: +0.90 балла в пользу группы с доп. занятиями

Статистическая проверка:

- t-test: p-value = 0.4180
- Тест Манна-Уитни: p-value = 0.2203

Вывод: Различия в успеваемости между группами **не являются статистически значимыми** ($p > 0.05$ в обоих тестах). Наблюдаемое преимущество группы с дополнительными занятиями (+0.90 балла) может быть случайным.

5. Анализ пропусков занятий

Корреляции:

- Пропуски и алкоголь: 0.139
- Пропуски и оценки: -0.006

Сравнение групп:

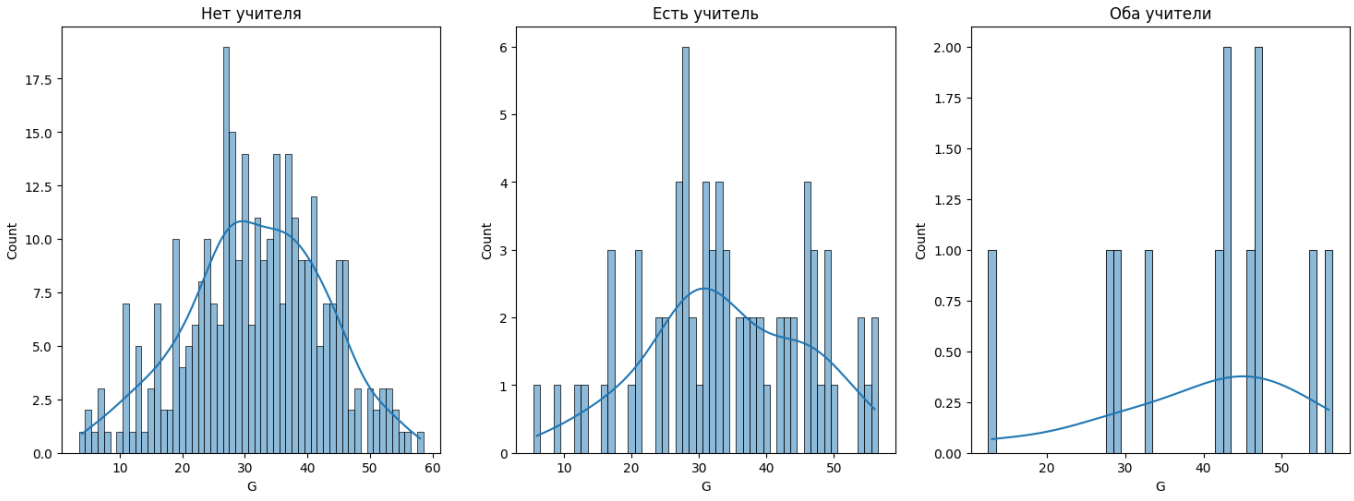
- Частые пропуски: средний Salc = 4.23
- Редкие пропуски: средний Salc = 3.49
- теста Манна-Уитни p-value = $0.0014 < 0.05$

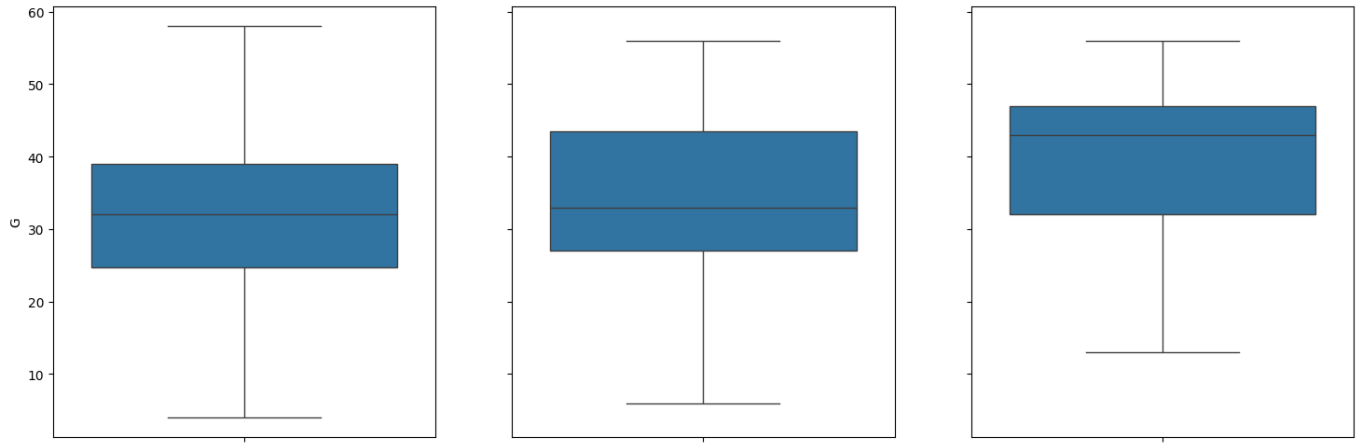
Вывод: Студенты с большим количеством пропусков потребляют больше алкоголя.

6. Семейные факторы

7. Может быть профессия учителя у родителей помогает?

Метод пристального взгляда:





Для первых двух случаев данные ещё как-то можно считать нормальными, поэтому проверим на двух тестах.

Без учитля vs с хотя бы одним учителем:

- Манна-Уитни: p-value = 0.06
- t-тест: p-value = 0.09

Нет оснований отвергнуть нулевую гипотезу о равенстве, следовательно, разница статистически не значима.

Без учителя vs два учителя:

Манна-Уитни: p-value = 0.004

С двумя учителями лучше.

Описание статистических тестов

Тест Манна-Уитни

- **Назначение:** Сравнение двух независимых выборок
- **Условия применения:** Отсутствие нормальности распределения, порядковые данные
- **Интерпретация:** Проверяет гипотезу о том, что одна выборка стахостически больше другой

t-тест для независимых выборок

- **Назначение:** Сравнение средних значений двух групп
- **Условия применения:** Нормальность распределения, гомогенность дисперсий
- **Интерпретация:** Проверяет гипотезу о равенстве средних

Тест Шапиро-Уилка

- **Назначение:** Проверка нормальности распределения
- **Условия применения:** Размер выборки < 5000
- **Интерпретация:** p-value > 0.05 означает отсутствие оснований отвергать нормальность

Тест D'Agostino

- **Назначение:** Проверка нормальности на основе асимметрии и эксцесса
- **Условия применения:** Подходит для больших выборок
- **Интерпретация:** Аналогично тесту Шапиро-Уилка

Основные выводы

1. **Статус родителей** не оказывает значимого влияния на потребление алкоголя студентами
2. **Мужчины** потребляют значительно больше алкоголя, чем женщины
3. **Дополнительные занятия** не улучшают успеваемость
4. **Пропуски занятий** сильно коррелируют с повышенным потреблением алкоголя
5. **Семейные факторы** демонстрируют слабые корреляции с алкоголем и успеваемостью

Наибольшее практическое значение имеет связь между пропусками занятий и потреблением алкоголя, что может быть использовано для разработки профилактических программ.