**Документация**

**Дипломная работа по теме:**

**“Анализ данных о курении и употреблении алкоголя (поиск инсайтов и составление рекомендаций стейкхолдерам, построение предиктивной модели)”**

**Профессия “Аналитик данных”, DAU-69**

**Сапрунов Александр Игоревич**

**г. Новосибирск, 2025**

**Оглавление**

[**Введение 3**](#_heading=h.gjdgxs)

[**Блок 1. Описание исходного датасета и типов данных 4**](#_heading=h.30j0zll)

[**Блок 2. Подготовка данных 5**](#_heading=h.1fob9te)

[**Блок 3. Аномалии и очистка данных 5**](#_heading=h.3znysh7)

[**Блок 4. Анализ данных для стейкхолдеров. 7**](#_heading=h.2et92p0)

[**Блок 5. Построение модели логистической регрессии 18**](#_heading=h.tyjcwt)

[**Итоги проекта и заключение 20**](#_heading=h.1t3h5sf)

# Введение

**Цели проекта:**

На основе собранных данных Национальной службы медицинского страхования в Кореи, понять какие признаки изменений в организме человека дают нам представление о том ведет человек здоровый образ жизни или склонен к увлечению плохими привычками (употребления алкоголя или табака), определить основных стейкхолдеров на основе выводов, полученных из проведенного анализа.

**Задачи проекта:**

1.Выявить значение медицинских показателей характеризующие образ жизни человека, а также корреляцию между этими показателями *(для трудоспособного населения).*

2.Построить предсказательную модель, чтобы выявить проблему там, где она уже есть.

**Стейкхолдеры:**

Органы государственной власти, медицинские учреждения, неправительственные организации (общественные организации).

# **Блок 1. Описание исходного датасета и типов данных**

Для исследования был взят датасет [Smoking and Drinking Dataset with body signal](https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset/data).

| № | Наименование столбца | Описание | Тип данных |
| --- | --- | --- | --- |
| 1 | sex | Пол (мужчина, женщина) | object |
| 2 | age | Возраст (округление до 5 лет) | int64 |
| 3 | height | Высота (округление до 5 см) | int64 |
| 4 | weight | Вес в кг | int64 |
| 5 | waistline | Объем талии в см | float64 |
| 6 | sight\_left | Зрение левый глаз | float64 |
| 7 | sight\_right | Зрение правый глаз | float64 |
| 8 | hear\_left | Слух слева (1-нормальный, 2 – отклонение) | float64 |
| 9 | hear\_right | Слух справа (1-нормальный, 2 – отклонение) | float64 |
| 10 | SBP | Систолическое артериальное давление [мм рт. ст.] | float64 |
| 11 | DBP | Диастолическое артериальное давление [мм рт. ст.] | float64 |
| 12 | BLDS | уровень глюкозы в крови натощак) [мг/дл] | float64 |
| 13 | tot\_chole | общий холестерин[мг/дл] | float64 |
| 14 | HDL\_chole | Холестерин ЛПВП [мг/дл] | float64 |
| 15 | LDL\_chole | Холестерин ЛПНП[мг/дл] | float64 |
| 16 | triglyceride | триглицериды[мг/дл] | float64 |
| 17 | hemoglobin | гемоглобин[г/дл] | float64 |
| 18 | urine\_protein | белок в моче, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4) | float64 |
| 19 | serum\_creatinine | Креатинин сыворотки (крови) [мг/дл] | float64 |
| 20 | SGOT\_AST | SGOT (глутамат-оксалоацетат трансаминаза) AST (аспартат трансаминаза) [МЕ/л] | float64 |
| 21 | SGOT\_ALT | АЛТ (Аланиновая трансаминаза) [МЕ/л] | float64 |
| 22 | gamma\_GTP | y-глутамилтранспептидаза [МЕ/л] | float64 |
| 23 | SMK\_stat\_type\_cd | Курение: 1 (никогда), 2 (курил, но бросил), 3 (все еще курю) | float64 |
| 24 | DRK\_YN | Пьющий-Y или нет-N | object |

# **Блок 2. Подготовка данных**

Согласно представленному датасету, в таблице присутствует 991346 строк, все данные имеют ненулевые значения. Прежде чем приступить к анализу необходимо проверить значение в каждом столбце с признаками на предмет недопустимых или аномальных значений, возникающие из ошибок, по возможности определим выбросы.

# **Блок 3. Аномалии и очистка данных**

При первичном анализе выявлены различные аномалии данных, несоответствующие или недопустимые значения, выбросы.

**3.1 Замена в данных**

Произведено приведение данных в величины, принятые в медицине в РФ.

| № | Наименование столбца | Тип данных | Произведенные действия с данными |
| --- | --- | --- | --- |
| 1 | sex | object | Заменены категории male и female на значения 1 и 0 соответственно |
| 24 | DRK\_YN | object | Заменены категории Y и N на значения 1 и 0 соответственно |
| 12 | BLDS | float64 | Переведены значения из mg/dl в mmol/l для соответствия привычным значениям |
| 13 | tot\_chole | float64 | Переведены значения из mg/dl в mmol/l для соответствия привычным значениям |
| 14 | HDL\_chole | float64 | Переведены значения из mg/dl в mmol/l для соответствия привычным значениям |
| 15 | LDL\_chole | float64 | Переведены значения из mg/dl в mmol/l для соответствия привычным значениям |
| 16 | triglyceride | float64 | Переведены значения из mg/dl в mmol/l для соответствия привычным значениям |
| 17 | hemoglobin | float64 | Переведены значения из g/dl в g/l для соответствия привычным значениям |
| 19 | serum\_creatinine | float64 | Переведены значения из mg/dl в mkmol/l для соответствия привычным значениям |

**3.2 Аномалии в данных**

Результат обработанного датасета, а также выводы по причинам очистки данных приведены в таблице:

| № | Наименование столбца | Тип данных | Причины изменения данных |
| --- | --- | --- | --- |
| 2 | age | int64 | Для своего проекта было принято решение оставить только трудоспособное население в возрасте для женщин от 25 до 60 лет, для мужчин от 25 до 65 лет. |
| 3 | height | int64 | Очистка не производилась |
| 4 | weight | int64 | Все строки со значением меньше 45 удалены и значением равным 999 (некорректные значения) |
| 5 | waistline | float64 | Все строки со значением меньше 54 удалены (выбросы) |
| 6 | sight\_left | float64 | Удалены строки со значением 9.9 (некорректные значения) |
| 7 | sight\_right | float64 | Удалены строки со значением 9.9 (некорректные значения) |
| 8 | hear\_left | float64 | Очистка не производилась |
| 9 | hear\_right | float64 | Очистка не производилась |
| 10 | SBP | float64 | Удалены строки, не попадающие в промежуток от 90 до 230 (некорректные значения) |
| 11 | DBP | float64 | Удалены строки, не попадающие в промежуток от 50 до 150 (некорректные значения) |
| 12 | BLDS | float64 | Удалены строки со значением больше 7.17 и со значением меньше 3.5 (выбросы) |
| 13 | tot\_chole | float64 | Удалены строки со значением больше 16.28 и со значением меньше 5.33 (выбросы) |
| 14 | HDL\_chole | float64 | Удалены строки со значением больше 5.39 и со значением меньше 0.94 (выбросы) |
| 15 | LDL\_chole | float64 | Удалены строки со значением больше 11.28 и со значением меньше 1.39 (выбросы) |
| 16 | triglyceride | float64 | Удалены строки со значением больше 16.22 (выбросы) |
| 17 | hemoglobin | float64 | Удалены строки со значением меньше 70 (выбросы) |
| 18 | urine\_protein | float64 | Очистка не производилась |
| 19 | serum\_creatinine | float64 | Удалены строки со значением больше 500 и со значением меньше 8.84 (выбросы) |
| 20 | SGOT\_AST | float64 | Удалены строки со значением больше 1000 (некорректные значения) |
| 21 | SGOT\_ALT | float64 | Удалены строки со значением больше 1000 (некорректные значения) |
| 22 | gamma\_GTP | float64 | Очистка не производилась |

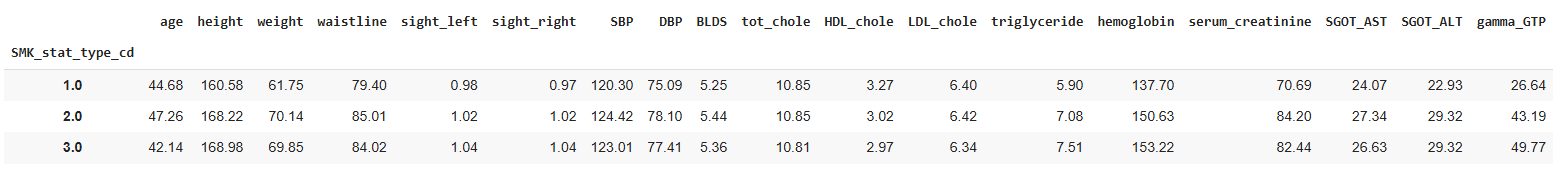
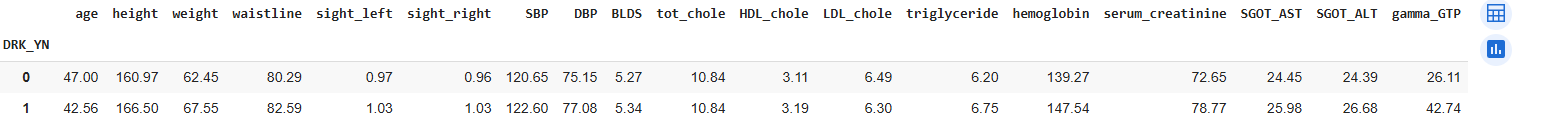
После удаления несоответствующих значений число строк стало равно 709014.

# **Блок 4. Анализ данных для стейкхолдеров.**

Целью блока является поиск тенденций и инсайтов для составления рекомендаций стейкхолдерам. Стейкхолдерами в нашем случае могут выступать государственные органы связанные с здравоохранением и социальным развитием.

**4.1. Анализ средних значений в целом по группам.**

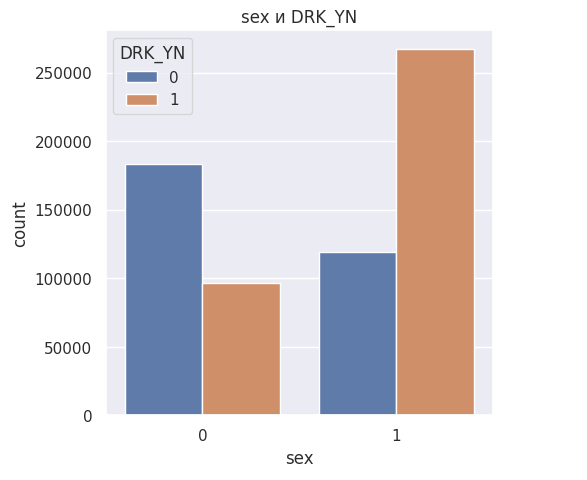
Группируемые признаки взяты без категориальных признаков.

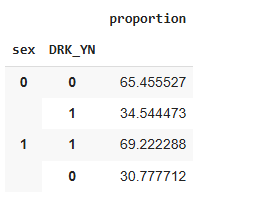


**Вывод по пункту 4.1**: хотя средние значения и отличаются, насколько они статистически значимы, требует дополнительной проверки.

При проведении группировки со средними значениями по отношению к алкоголю и курения выявлены небольшие отклонения средних значений в большую сторону по группам пациентов.

**4.2. Анализ категориальных значений и отношение к алкоголю.**

4.2.1 Гендерное сравнение отношения к алкоголю.

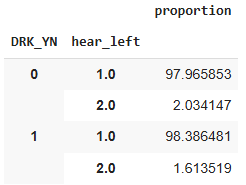
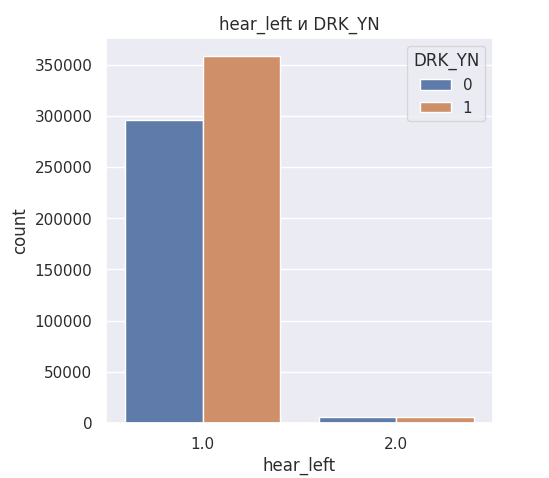


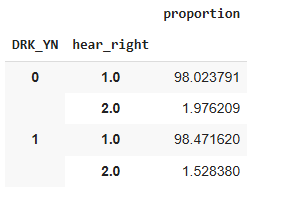
В наблюдаемом наборе данных выделяются следующие зависимости:

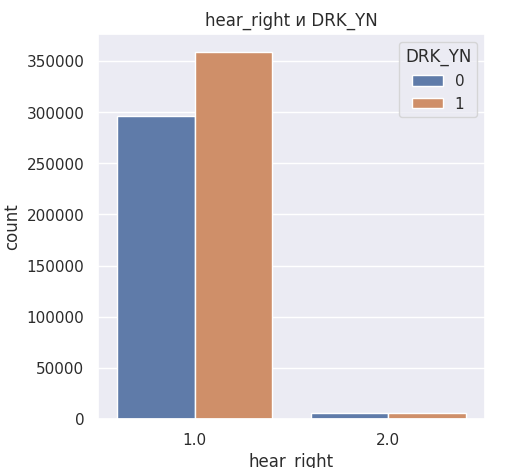
* 2/3 мужчин выпивают и наоборот 2/3 женщин не выпивают.

При подобном исследовании степень употребления алкоголя выяснить невозможно.

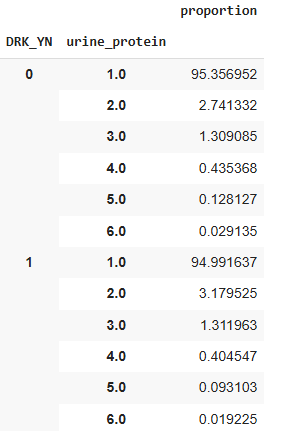
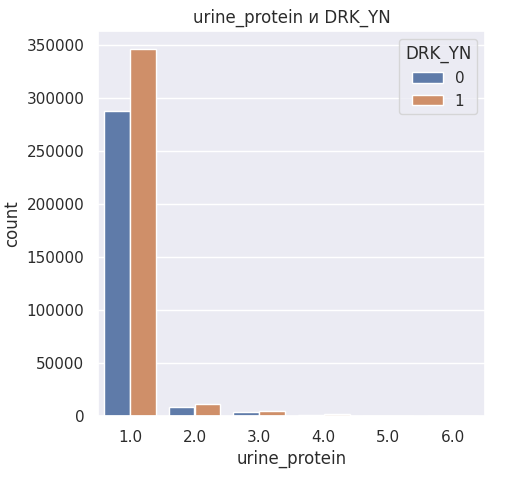
4.2.2. Степень нарушения слуха и употребление алкоголя для левого и правого уха.





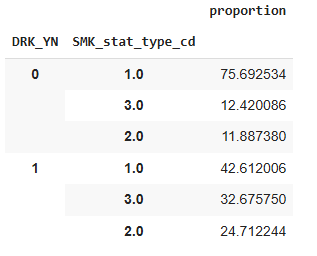


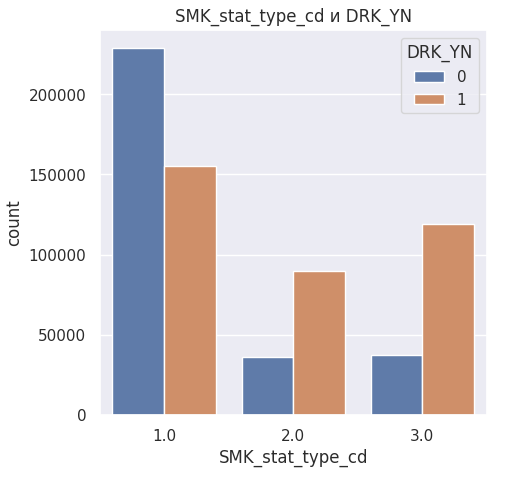
Хороший слух имеет 98% от числа наблюдаемых, не зависит от употребления алкоголя.

4.2.3 Сравнение нахождения белка в моче и употребление алкоголя.

Степень наличия белка в моче не указывает на факт употребления алкоголя.

4.2.4 Сравнение курения и употребления алкоголя.

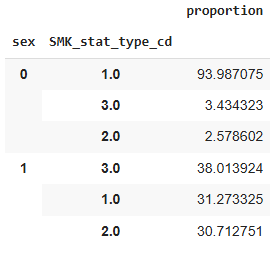
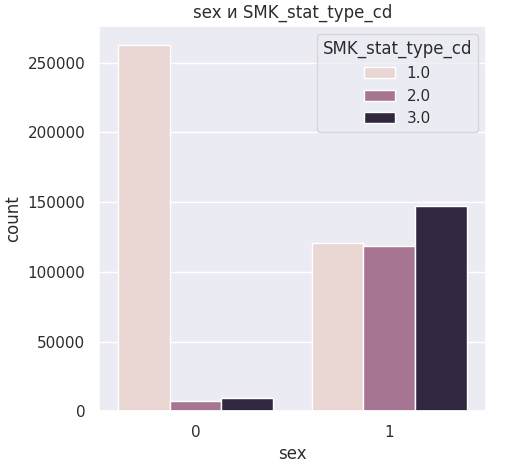




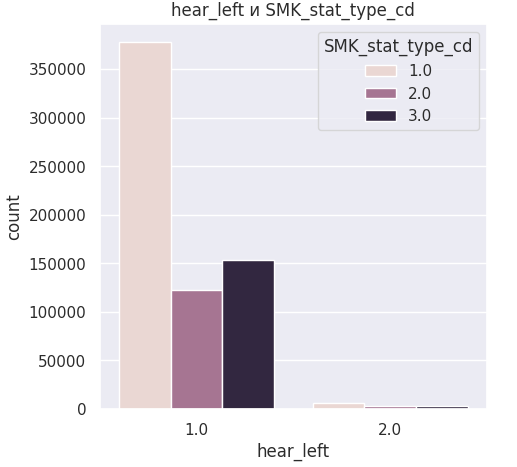
Пациенты, проявляющие толерантность к курению - 76% не курящих и также проявляют независимость к алкоголю - 43% не пьющих, далее проявляется обратная зависимость, те кто курил или курит склонны больше употреблять алкоголь.

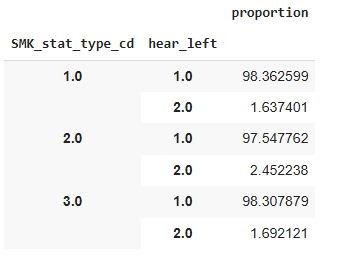
**4.3. Анализ категориальных значений и курение. (1-не курит, 2-бросил, 3-курит).**

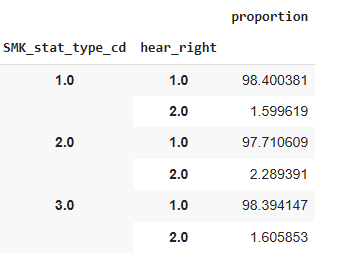
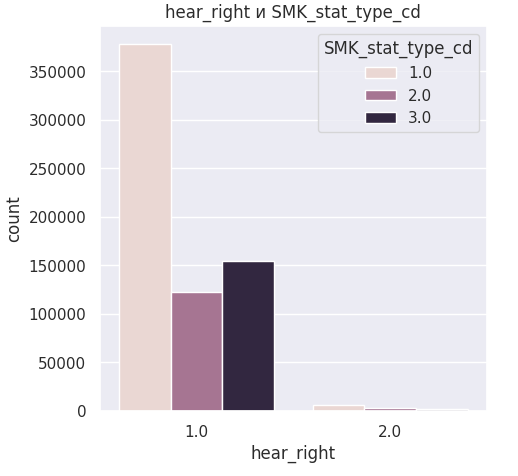
4.3.1 Гендерное сравнение отношения к алкоголю.



94 % женщин не курят, мужчины же разделились практически в равных пропорциях.

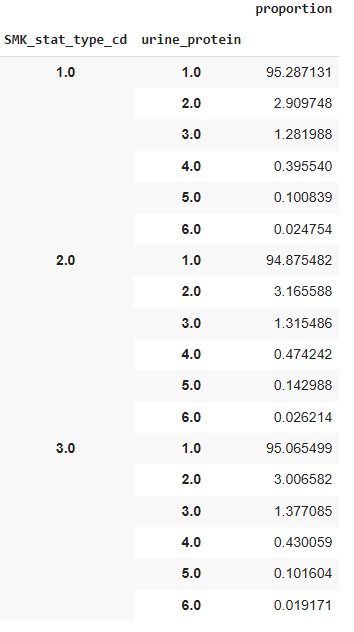
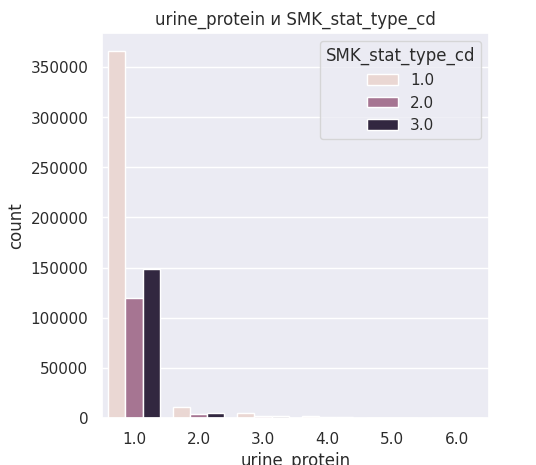
4.3.2. Степень зависимости нарушения слуха и курения для левого и правого уха.





Курение не влияет на слух.

4.3.3 Сравнение нахождения белка в моче и употребление алкоголя.



Степень наличия белка в моче не указывает на факт курения.

**Вывод:** по пункту 4.2 и 4.3.

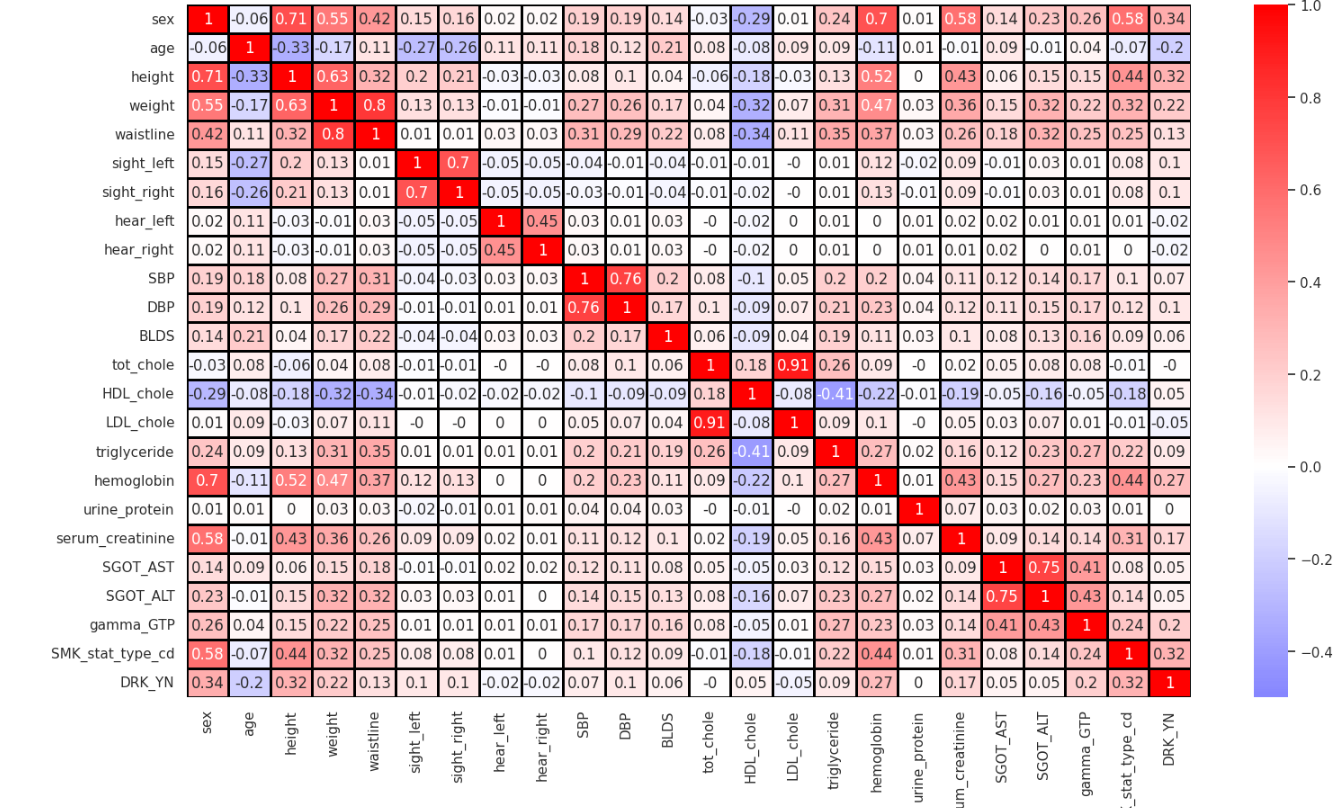
По анализу категориальных значений можно сделать следующий вывод, что доля мужчин как употребляющих алкоголь, так и курящих выше, чем у женщин. Это говорит нам о том, что здесь имеется какая-то социальная составляющая поскольку мужское общество не здорово и с пренебрежением относится к своему здоровью.

**Рекомендации:**

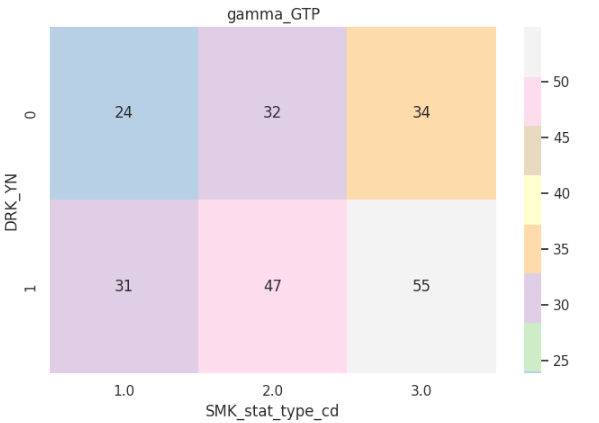
Нужно государственным органам акцентировать на этом внимание, усилить пропаганду здорового образа жизни среди мужского населения.

**4.4. Многофакторный анализ.**

4.4.1. Тепловая карта корреляции на основе очищенных данных.

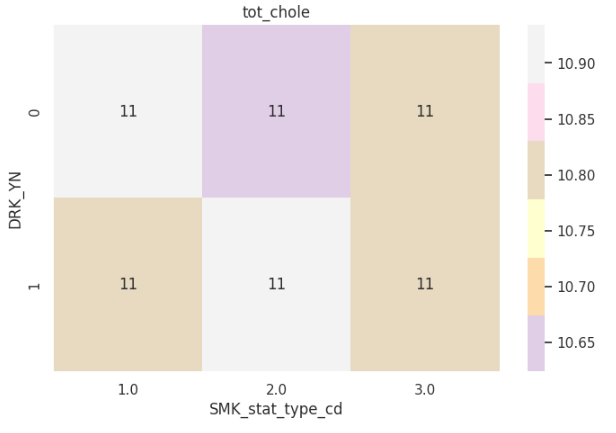


Вывод: согласно представленной таблице, мы можем увидеть некоторые тенденции, есть показатели, которые коррелируют только с родственными ему показателями, например рост и вес, вес и объем талии, правый и левый глаз, правое и левое ухо, систолическое и диастолическое давление, показатели холестерина между собой, а также SGOT\_AST и SGOT\_ALT. С контрольными показателями прослеживаются очень слабые взаимозависимости.

4.4.2 Матрица средних значений отношения к курению и алкоголю и глутамилтранспептидаза (*gamma\_GTP*).

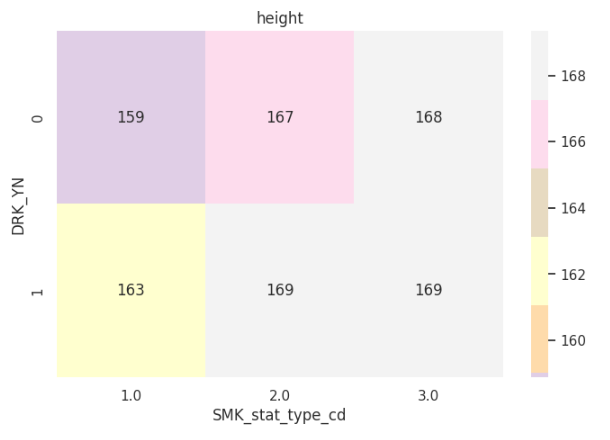
Вывод: видно, что средние значения показателя растут, от здорового образа жизни к противоположному концу, где пациенты и курят и выпивают.

4.4.3 Матрица средних значений отношения к курению и алкоголю и общий холестерин (*tot\_chole*).

Вывод: показатель не показывает значительной разницы, что у здоровых, что и у пациентов с зависимостью.

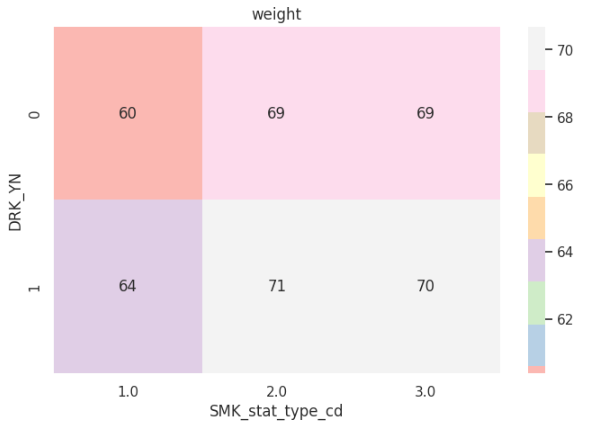
Вывод: на основе пунктов 4.4.2 и 4.4.3 проверять значения с контрольными показателями есть смысл там, где корреляция имеет значения от 0,2 ед.

4.4.4 Матрица средних значений отношения к курению и алкоголю и рост (*height*).



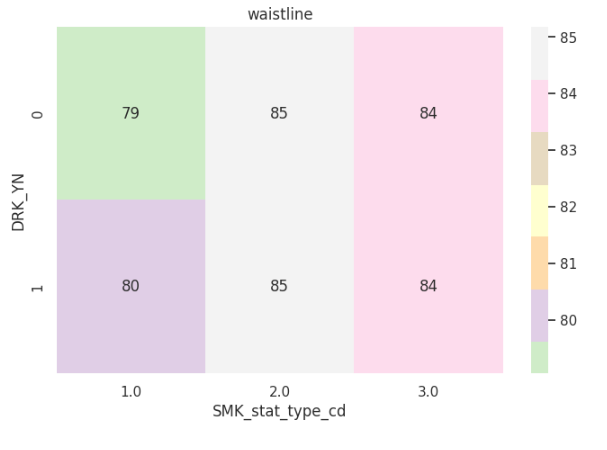
Вывод: Средние значения показателя растут, от здорового образа жизни к противоположному концу, где пациенты и курят и выпивают.

4.4.5 Матрица средних значений отношения к курению и алкоголю и вес (*weight*).



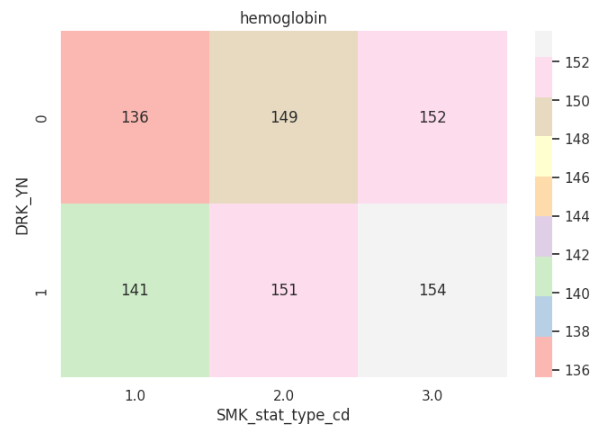
Вывод: Средние значения показателя растут, от здорового образа жизни к противоположному концу, где пациенты и курят и выпивают.

4.4.6 Матрица средних значений отношения к курению и алкоголю и объем талии (*waistline*)

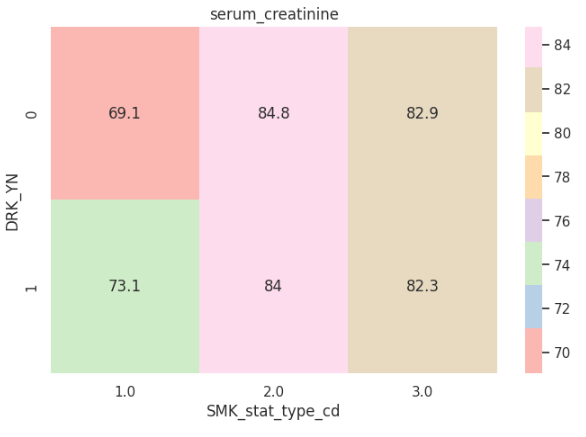


Вывод: Средние значения показателя растут, от здорового образа жизни к противоположному концу, где пациенты и курят и выпивают.

4.4.7 Матрица средних значений отношения к курению и алкоголю и гемоглобин (*hemoglobin*).

Вывод: Средние значения показателя растут, от здорового образа жизни к противоположному концу, где пациенты и курят и выпивают.

4.4.9 Матрица средних значений отношения к курению и алкоголю и Креатинин сыворотки (крови) (serum\_creatinine).



Вывод: Средние значения показывают примерно одинаковые результаты в каждой группе, зависимость показателей есть, но она не линейная.

**Вывод** на основе проведенного анализа:

можно выдвинуть гипотезу (предположение), что есть показатели, которые дают нам положительную корреляцию с отношением к алкоголю и курению, а чаще, когда два этих фактора связаны между собой это видно при сравнении показателей в верхнем левом углу и нижнем правом углу.

**Рекомендации:**

При проведении медицинских комиссий, диспансеризации медицинским организациям обращать внимание на признаки указывающие, что пациенты употребляют алкоголь или никотин, доносить до пациентов о недопустимости такого образа жизни.

# **Блок 5. Построение модели логистической регрессии**

# **Логистическая регрессия (ЛогР)** — статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события, она выдает ответ в виде числа в промежутке от 0 до 1. Бинарная логистическая регрессия применяется в случае, когда зависимая переменная является бинарной (может принимать только два значения). С помощью ЛогР можно оценивать вероятность наступления события.

В отличие от обычной регрессии, в методе логистической регрессии не производится предсказание значения числовой переменной исходя из выборки исходных значений. Вместо этого, значением функции является вероятность того, что данное исходное значение принадлежит к определенному классу.

**5.1 Подготовка данных**

Целью блока является построение модели логистической регрессии для выявления целевого показателя Пьющий-Y или нет-N.

Условия, которым должны удовлетворять данные:

| Условие | Соответствие данных |
| --- | --- |
| Наблюдения (входные данные) должны быть независимы друг от друга. | Используемые признаки - age, height, weight, sight\_left, hear\_left, SBP, BLDS, tot\_chole, HDL\_chole, triglyceride, hemoglobin, urine\_protein, SGOT\_AST, gamma\_GTP, SMK\_stat\_type\_cd. |
| Между признаками (независимыми входными переменными) не должно быть мультиколлинеарности \*. | Удалены столбцы - sex, waistline, sight\_right, DBP, LDL\_chole, SGOT\_ALT. |
| Зависимая переменная должна иметь бинарное значение, то есть 0 или 1 . | Соответствует, по столбцу DRK\_YN данные приведены к значениям 0 и 1. |
| Достаточное количество обучающих примеров . | Соответствует 532923- строк и 17 столбцов. |
| В обучающих данных не должно быть много выбросов . | Проделана работа по удалению выбросов и несуществующих значений. |

\* Две переменных коллинеарны, когда они находятся между собой в линейной зависимости, если коэффициент корреляции более 0,7. Чтобы избавиться от мультиколлинеарности, необходимо исключить из модели один из факторов.

**5.2 Оценка качества модели**

**Точность (accuracy)** - представляет собой долю правильно классифицированных объектов из общего числа объектов в наборе данных

Точность (accuracy) на тренировочной части - 0,699

Точность (accuracy) на тестовой части - 0,70

**Точность (precision)** **-**точность показывает долю объектов, названных моделью положительными, которые действительно являются положительными **— 0,73**

**Полнота (recall) или чувствительность -**полнота показывает долю объектов положительного класса, которые были правильно выявлены моделью **- 0,73**

**5.3 Кроссвалидация**

Это метод оценки качества модели машинного обучения, который помогает определить, насколько хорошо модель будет работать на новых, невиданных данных.

**K-блочная кросс-валидация - 0,69**

**Вывод:** Точность построенной модели не позволяет нам 100% выявлять некоторые особенности пациентов, но дает направление для поиска причины отклонения показателей здоровья от нормальных.

Созданная предиктивная модель позволяет прогнозировать вероятность начала курения и употребления алкоголя на основе ряда факторов. Это может быть полезно для разработки целенаправленных интервенций и профилактических программ.

# Итоги проекта и заключение

В рамках проекта был проведен анализ показателей пациентов и их взаимосвязей, были выявлены некоторые закономерности между этими показателями, а также сделаны выводы о том, что не все показатели могут быть связаны между собой, иные показатели проявлялись во всех группах одинаково, то есть имели свою тенденцию распределения. Попробовали построить модель, основанную на показателях здоровья для получения прогноза определения состояния человека.

Кроме того, исследование показало, что существует прямая корреляция между частотой курения и употребления алкоголя. Это подчеркивает необходимость более глубокого изучения психоэстетических факторов, способствующих формированию вредных привычек.

**Рекомендации:**

На основе полученных данных разработаны рекомендации для различных стейкхолдеров, включая:

- Региональные органы здравоохранения – усилить профилактические меры в группах высокого риска;

-Образовательные учреждения – внедрять программы по повышению осведомленности о последствиях курения и алкоголизма;

-Неправительственные организации – организовывать курсы и семинары, способствующие здоровому образу жизни.

Наблюдение за здоровьем населения является важной частью заботы государства о гражданах, такие исследования помогают понять не только состояние человеческого здоровья, а также состояние социума в целом, искать не только выводы, но и причины такого социального антиповедения, как курение и алкоголизм. Крепкое здоровье население это сила и опора для страны.

**Перспективы дальнейших исследований:** рекомендуется провести дальнейшие исследования на более широких выборках с использованием различных методов сбора данных, что позволит глубже понять динамику потребления табака и алкоголя, а также оценить эффективность предложенных мероприятий.