Анализ данных сердечных заболеваний

Bas Albert­­1

1-Технический университет Молдовы

Email: [albert.bas@iis.utm.md](mailto:albert.bas@iis.utm.md)

**Абстракт**

Сердечно-сосудистые заболевания (ССЗ) являются глобальной угрозой общественному здравоохранению, представляя собой ведущую причину смертности по всему миру. Оценка и анализ этих заболеваний играют ключевую роль в поддержании общественного здоровья и предупреждении осложнений. Различные факторы, такие как генетика, образ жизни и окружающая среда, могут способствовать развитию сердечных заболеваний.

Анализ данных о ССЗ становится критически важным инструментом для выявления рисков, разработки эффективных методов диагностики и лечения, а также для формирования стратегий предупреждения. В статье рассматривается значимость современных методов исследований, включая применение новых технологий в области медицинской диагностики и обработки данных, для обеспечения более точного анализа и эффективного вмешательства.

Подчеркивается важность использования датасетов, таких как UCI Heart Disease Data, в создании моделей для прогнозирования риска заболевания. Рассматриваются ключевые параметры этого набора данных, включая возраст, пол, характер боли в груди, артериальное давление и другие, для выявления связей и формирования практических выводов.

Исследование подчеркивает, что точное и своевременное выявление факторов риска, а также персонализированный подход к лечению, могут существенно снизить заболеваемость и смертность от ССЗ. Освещается необходимость дальнейших исследований и применения новейших технологий для более глубокого понимания и эффективного управления сердечно-сосудистыми заболеваниями.

**Введение**



Сердечно-сосудистые заболевания являются одной из ведущих причин смертности по всему миру, создавая серьезные вызовы для общественного здравоохранения. Оценка и анализ этих заболеваний играют важную роль в поддержании общественного здоровья и предупреждении возможных осложнений. Сердечные заболевания, такие как ишемическая болезнь сердца, артериальная гипертензия и сердечная недостаточность, могут иметь различные причины, включая генетические факторы, образ жизни, а также окружающую среду.

Анализ данных о сердечных заболеваниях становится важным инструментом для выявления рисковых факторов, разработки эффективных методов диагностики и лечения, а также для формирования стратегий предупреждения. Точное и своевременное выявление факторов риска и разработка персонализированных подходов к лечению могут значительно снизить заболеваемость и смертность от сердечно-сосудистых заболеваний.

В данном контексте анализ сердечных заболеваний не только способствует более раннему выявлению патологий, но и позволяет разрабатывать более эффективные стратегии профилактики, направленные на улучшение общественного здоровья. Этот процесс требует современных методов исследований, включая использование новых технологий в области медицинской диагностики и обработки данных, чтобы обеспечить более точный анализ и более эффективное вмешательство.

Изучение датасетов по заболеваниям сердца может помочь в создании моделей для прогнозирования риска заболевания, что в свою очередь может улучшить диагностику и предупреждение.

UCI Heart Disease Data

Этот набор данных представляет собой многомерный набор информации, включающий в себя множество отдельных математических или статистических переменных и предназначенный для многомерного анализа числовых данных. В нем содержится 14 атрибутов, таких как возраст, пол, характер боли в груди, артериальное давление в покое, уровень холестерина и сахара в крови, результаты электрокардиографии в покое, максимальная частота сердечных сокращений, стенокардия, вызванная физической нагрузкой, депрессия ST (oldpeak), наклон пикового сегмента ST при физической нагрузке, количество крупных сосудов и талассемия. Несмотря на наличие 76 признаков в общей базе данных, все проведенные исследования фокусировались на использовании подмножества из 14 признаков. Кливлендская база данных является единственной, которую исследователи в области машинного обучения применяли до настоящего времени. Главная цель этой базы данных заключается в прогнозировании, основываясь на предоставленных атрибутах пациента, наличия у него сердечного заболевания. Кроме того, она служит экспериментальной задачей для диагностики и извлечения информации, способной более глубоко осветить проблематику данного медицинского состояния.

1. id (Уникальный идентификатор для каждого пациента)
2. age (Возраст пациента в годах)
3. origin (Место обучения)
4. sex (Пол: Мужской/Женский)
5. cp (Тип боли в груди: [типичная стенокардия, атипичная стенокардия, нестенокардия, бессимптомная])
6. trestbps (Артериальное давление в покое при поступлении в больницу, в мм рт. ст.)
7. chol (Сывороточный холестерин в мг/дл)
8. fbs (Наличие гипергликемии: если уровень сахара в крови натощак > 120 мг/дл)
9. restecg (Результаты электрокардиографии в покое)
10. thalach (Максимальная достигнутая частота сердечных сокращений)
11. exang (Стенокардия, вызванная физической нагрузкой: Да/Нет)
12. oldpeak (Депрессия сегмента ST, вызванная физической нагрузкой относительно покоя)
13. slope (Наклон пика сегмента ST при максимальной физической нагрузке)
14. ca (Количество крупных сосудов, окрашенных при флюороскопии, от 0 до 3)
15. thal (Состояние талассемии: [нормальное; фиксированный дефект; обратимый дефект])
16. num (Предсказанный атрибут: наличие сердечного заболевания, категория)
17. num\_bi – созданная переменная, которая демонтирует лишь наличие или отсутствие болезни, без стадий

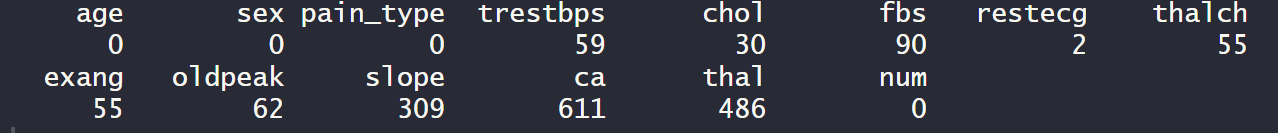
Количество элементов 930

**Анализ данных и создание модели**

Загрузка и очистка данных:

Мною были загружены необходимые библиотеки и данные сначало с сайта kaggle, затем в среде R

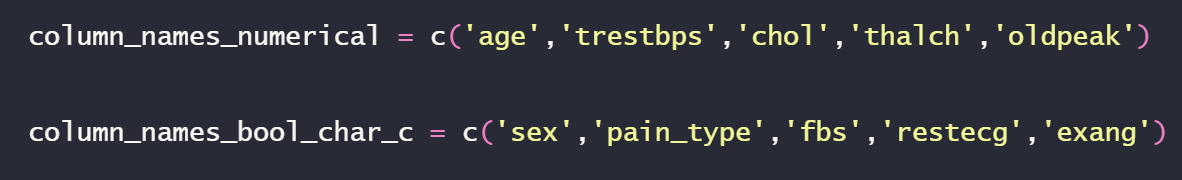
Первичные анализ None и NaN элементов показал, что некоторые колонки практически не имеют значений



Изображений 1. Количество nan значений по колонкам

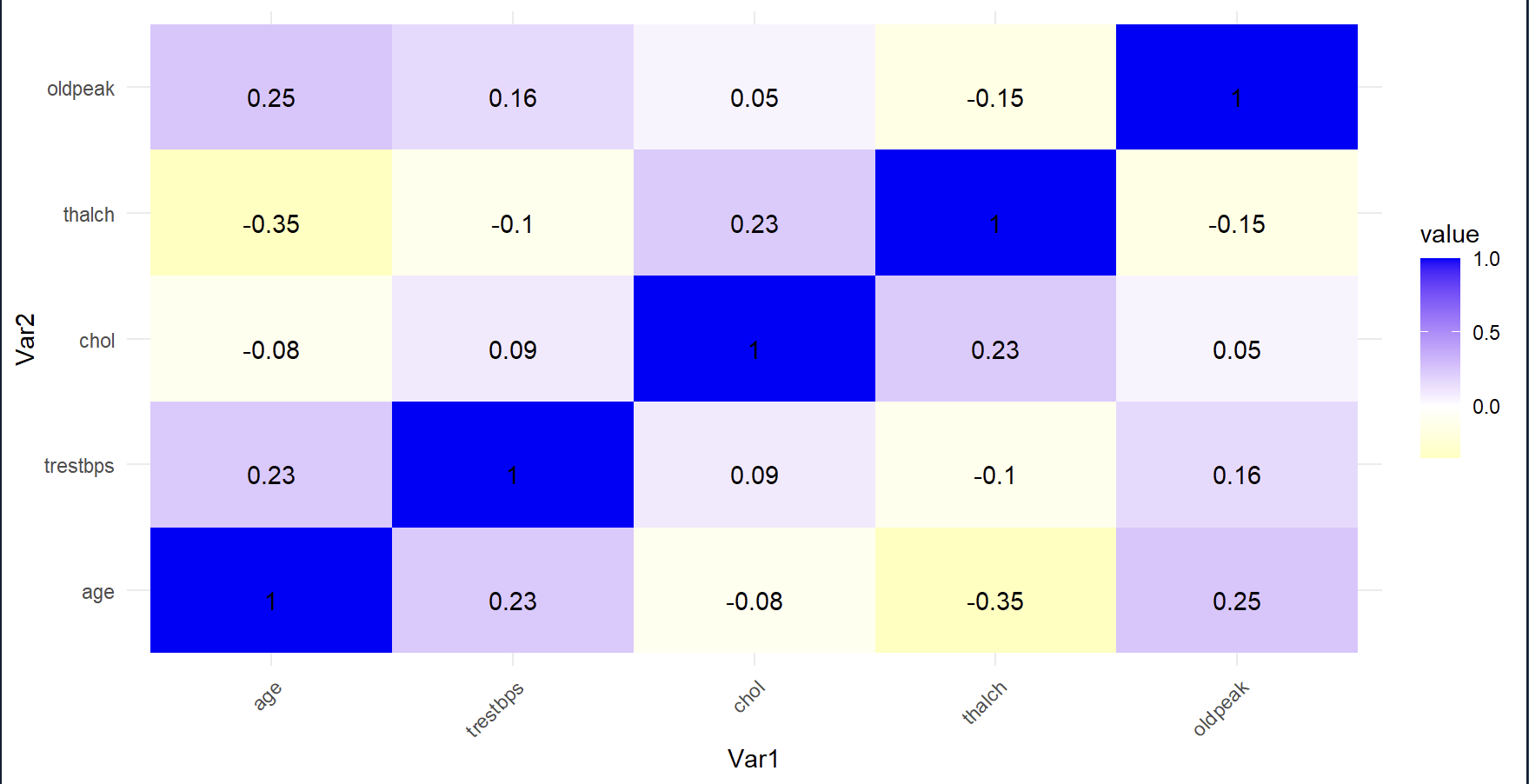
Переменные Slope ca thal имеют слишком мало значений, чтобы по ним делать какие либо выводы, больше, поэтому эти колонки удалены, чтобы уменшить риск их мнимого влияния на результаты.

Далее мною было выявлено какие колонки к каким типам данным относятся



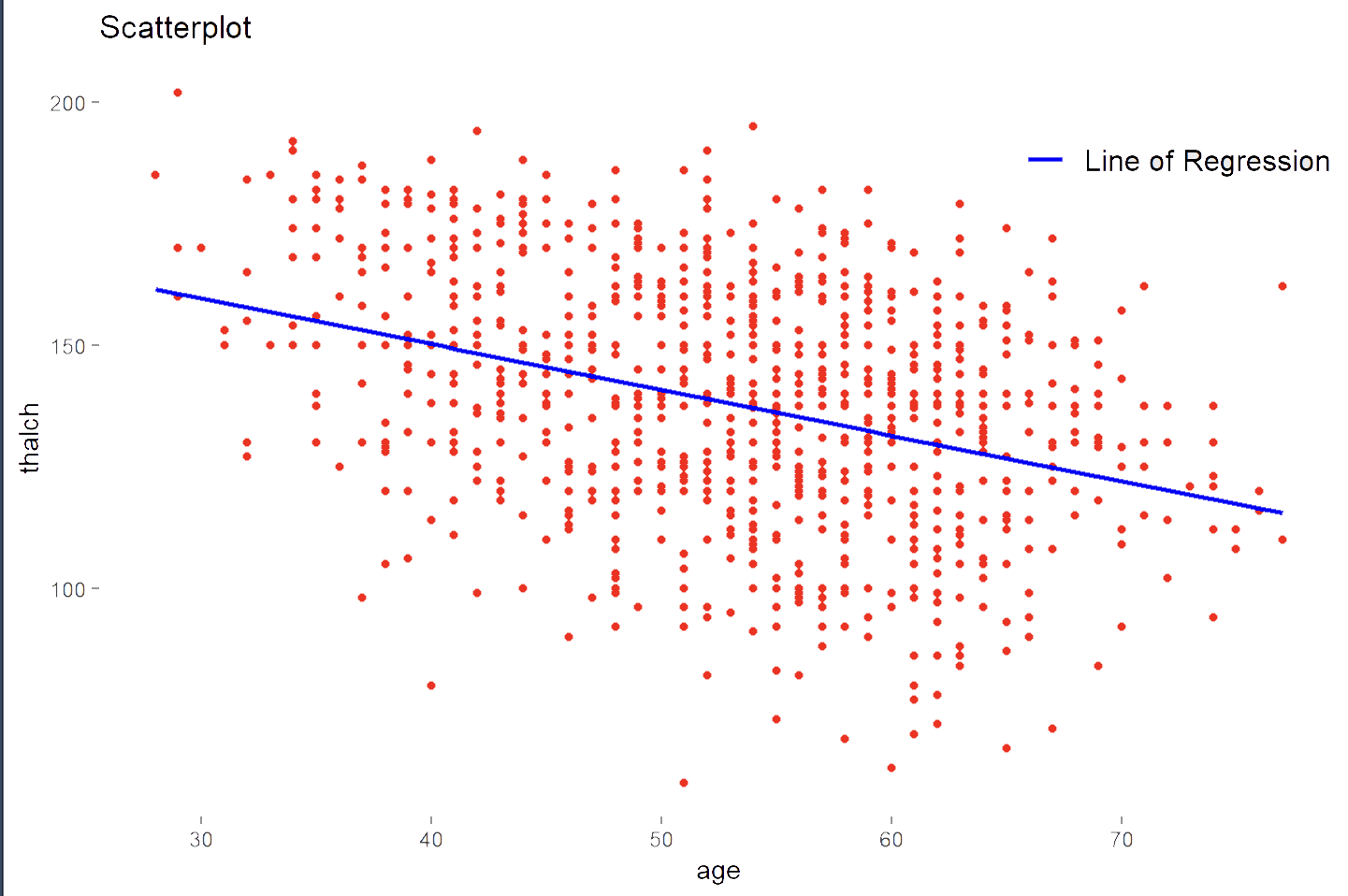
Изображений 2. Названия колонок по их типам данных

Далее nan значения в колонках непрерывных переменных было заменено на среднее значение, а категориальных – модой. Это снизит риск порчи результатов.



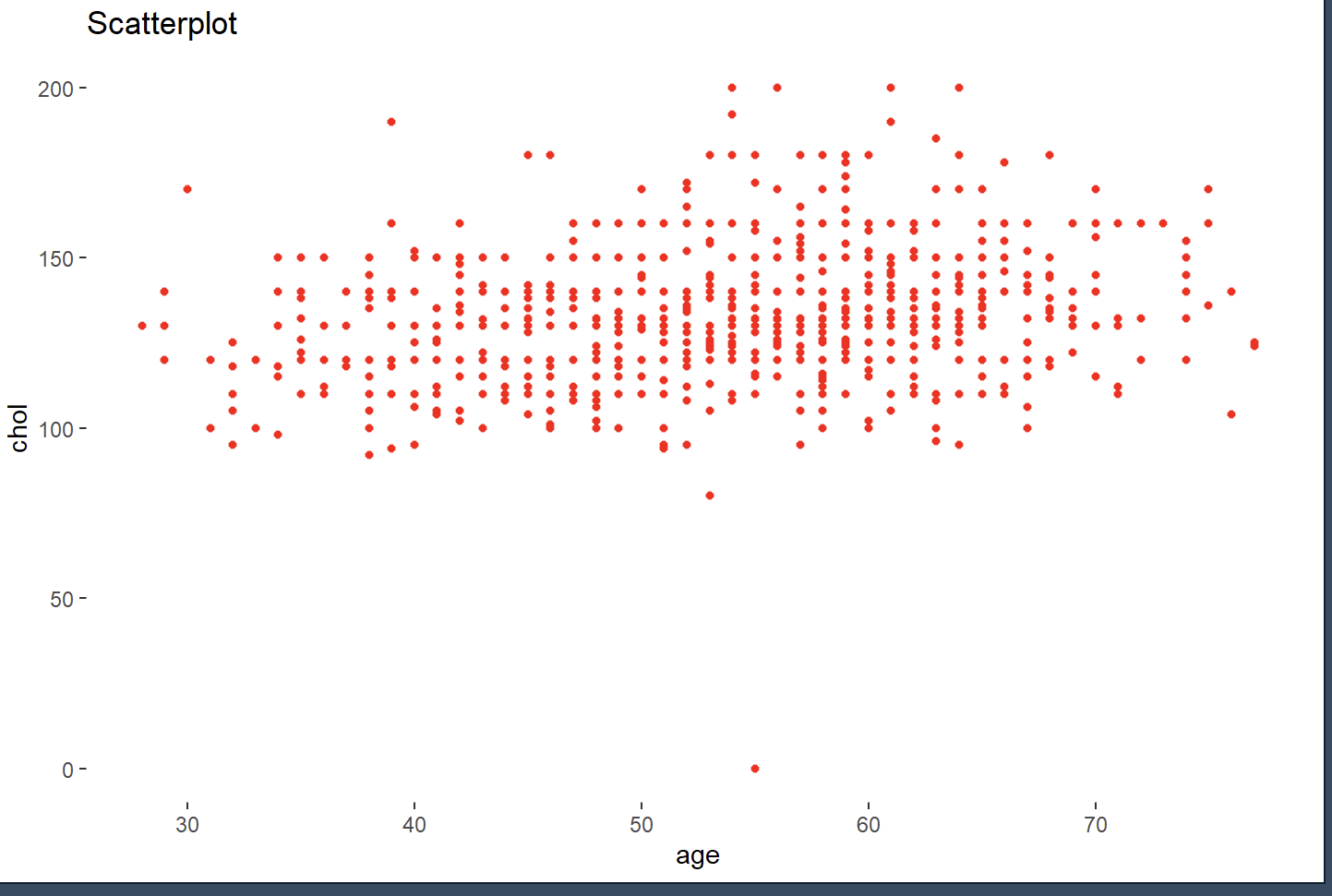
Изображение 3. Матрица корреляций

По графику корреляций видно, что она отсутствует, лишь небольшая взаимосвязь thalch и age



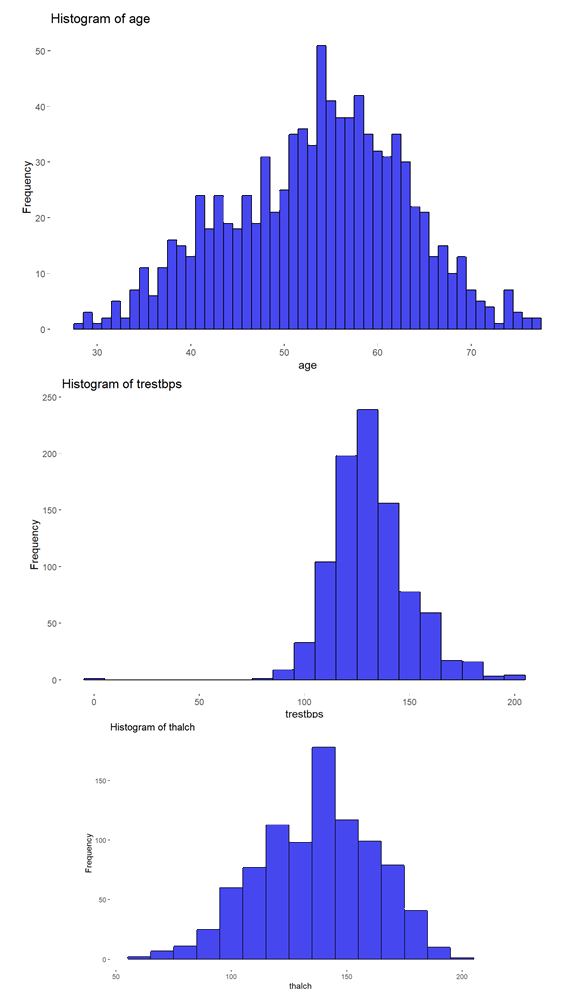
Изображение 4. Небольшая отрицательная корреляция

У других переменных корреляция отсутствует.

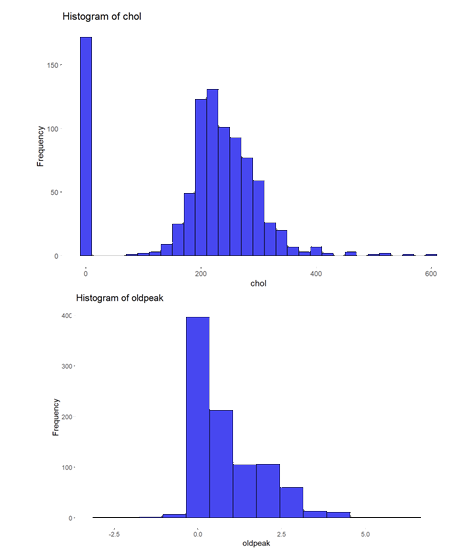


Изображение 5. Отсутствующая корреляция age и chol

Данные необходимо проверить на распределения, чтобы выявить аномальные значения ,графики и поведение переменных

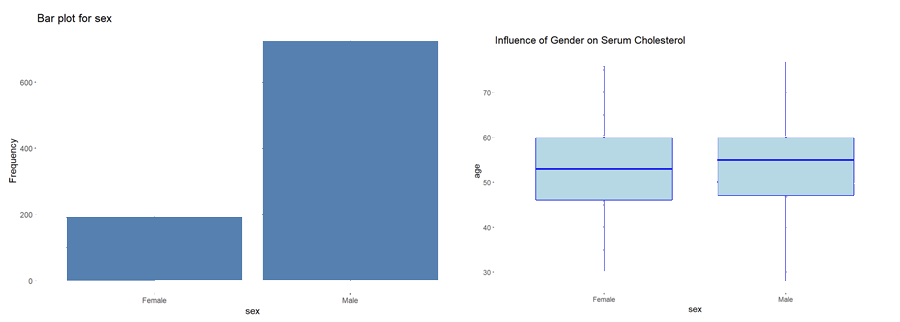


Изображение 6. Близкое к нормальному распределение thalch,Trestbps и age



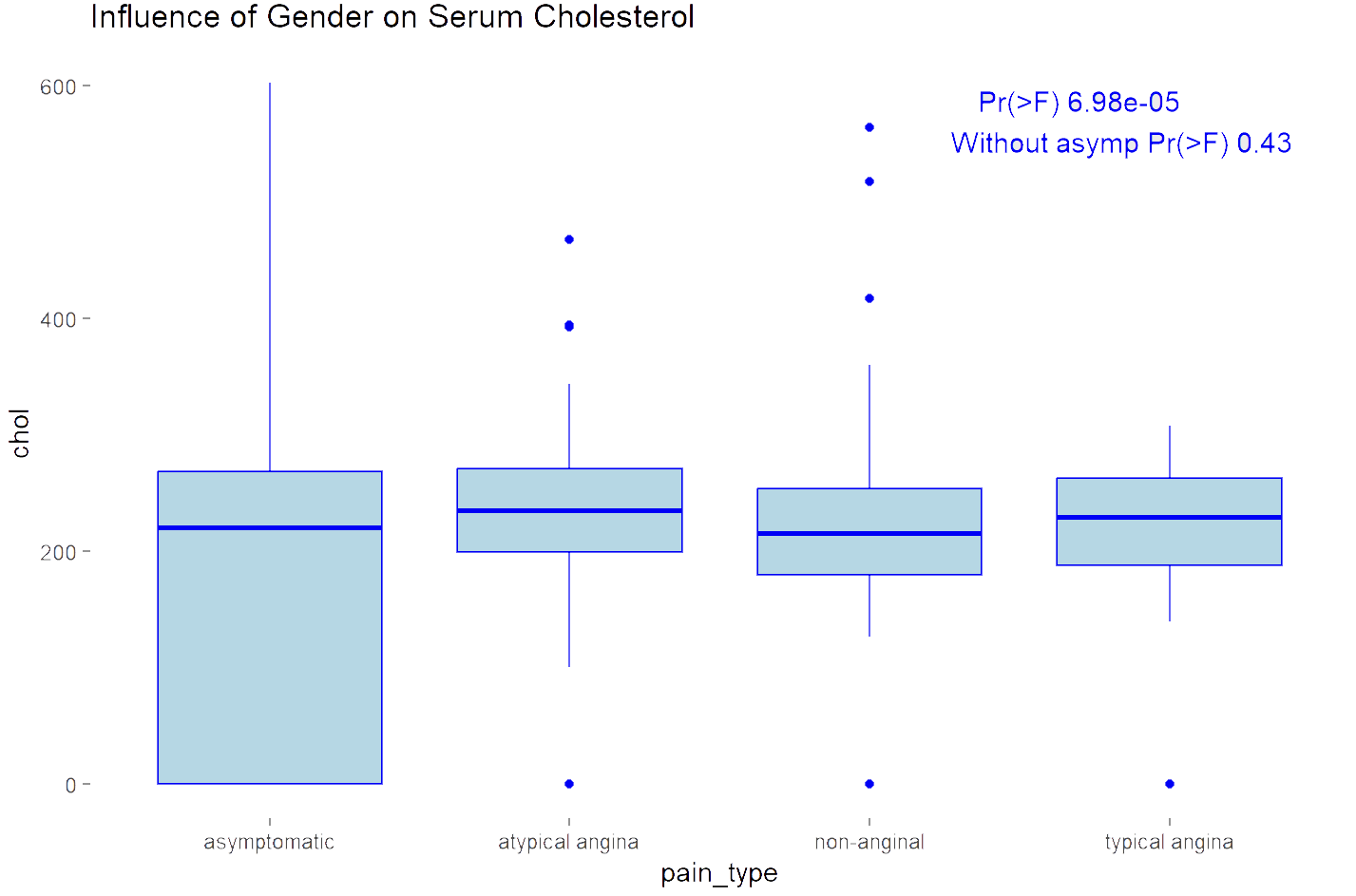
Изображение 7. Распределния oldpeak и chol

В датасете преобладает мужской пол, однако влияение этого фактора на распределения переменных выявлено не было



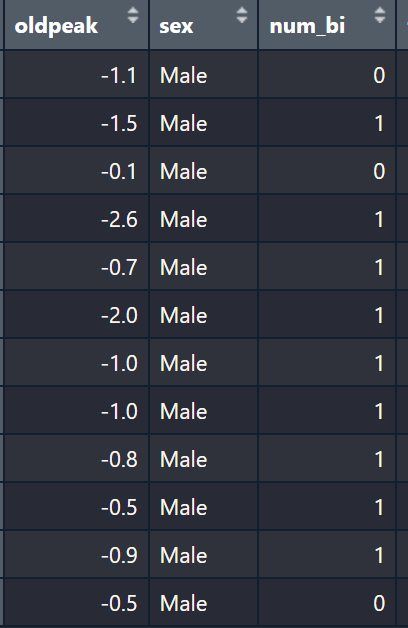
Изображение 8. Соотношение пола и возраста

Тип болезни, а именно асимптоматический сильно взаимодействует с показателем chol, при данной категории значение стемится к нулю , статистический тест дисперсии показал значимое влияние лишь при начилии данной категории и выбросов



Изображение 8.1 Chol и pain-type box-plot

Показатель olpeak снижался ниже 0 лишь у мужского пола, данные пациенты в большей своей части болели сердечным заболеванием

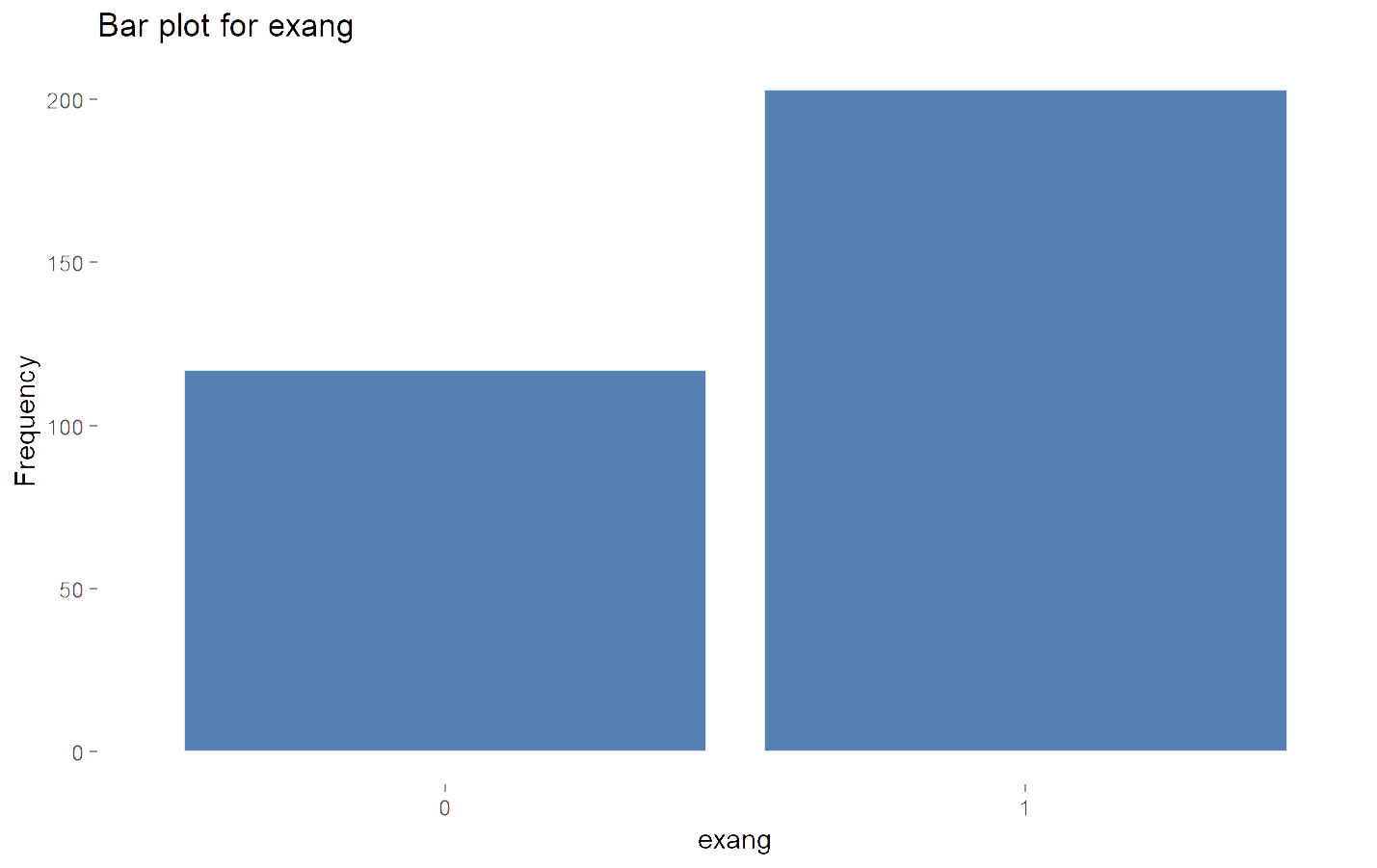
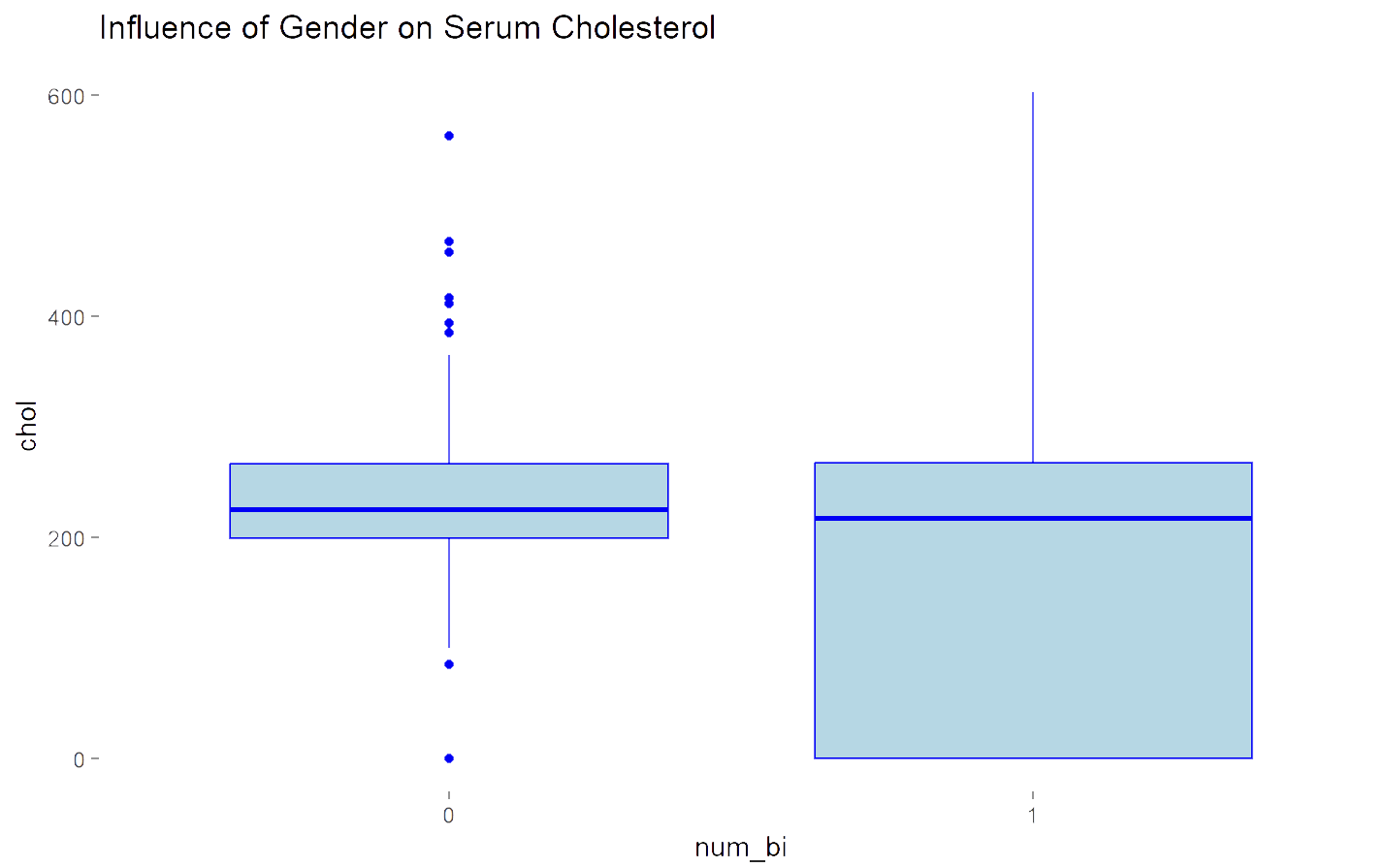




Изображение 8.2 Влияние olpeak

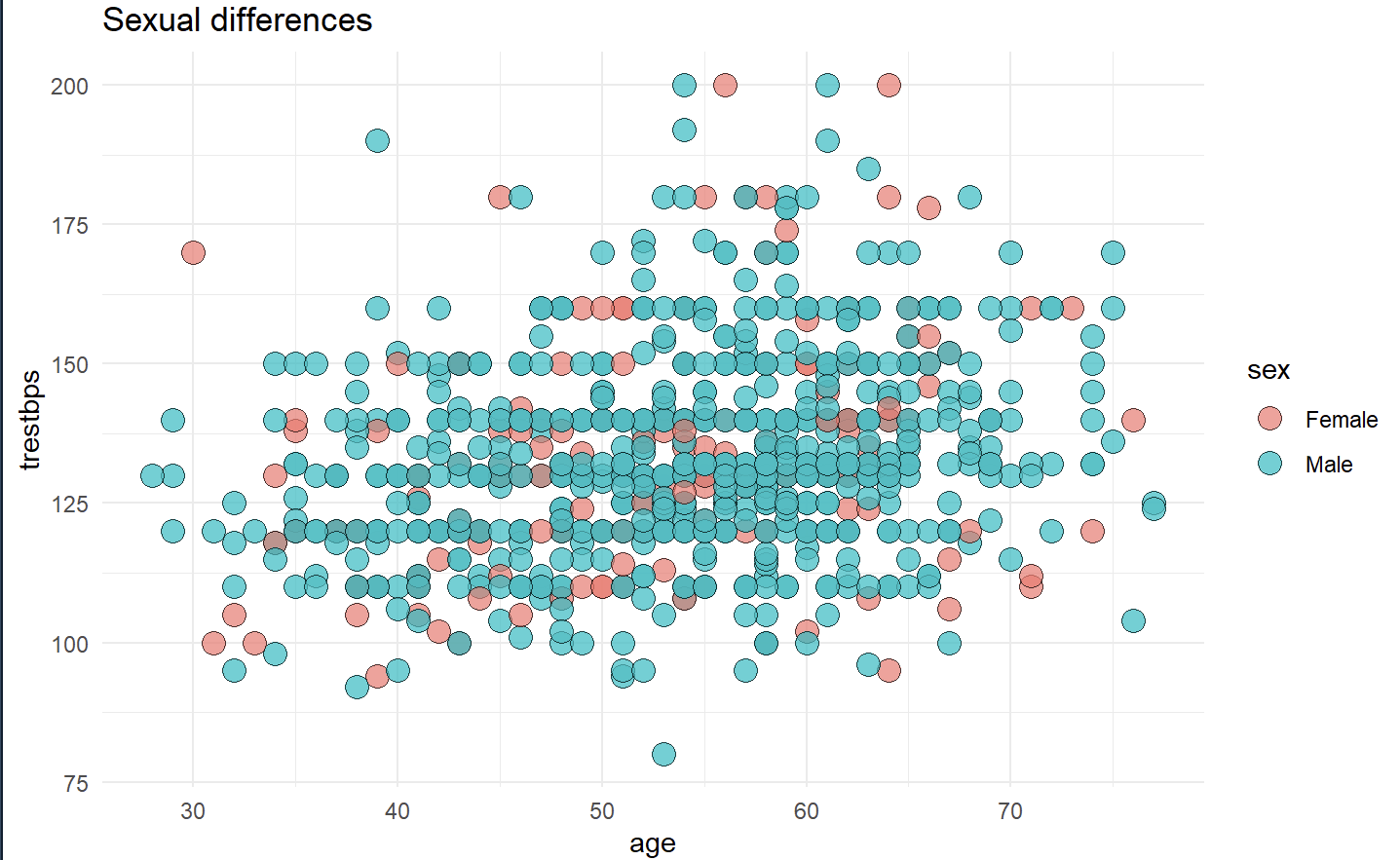
Было выявлено, что у большинства людей у кого коказатель chol ниже 200 болеют

Изображение 9. Box-plot chol для категорий болезни

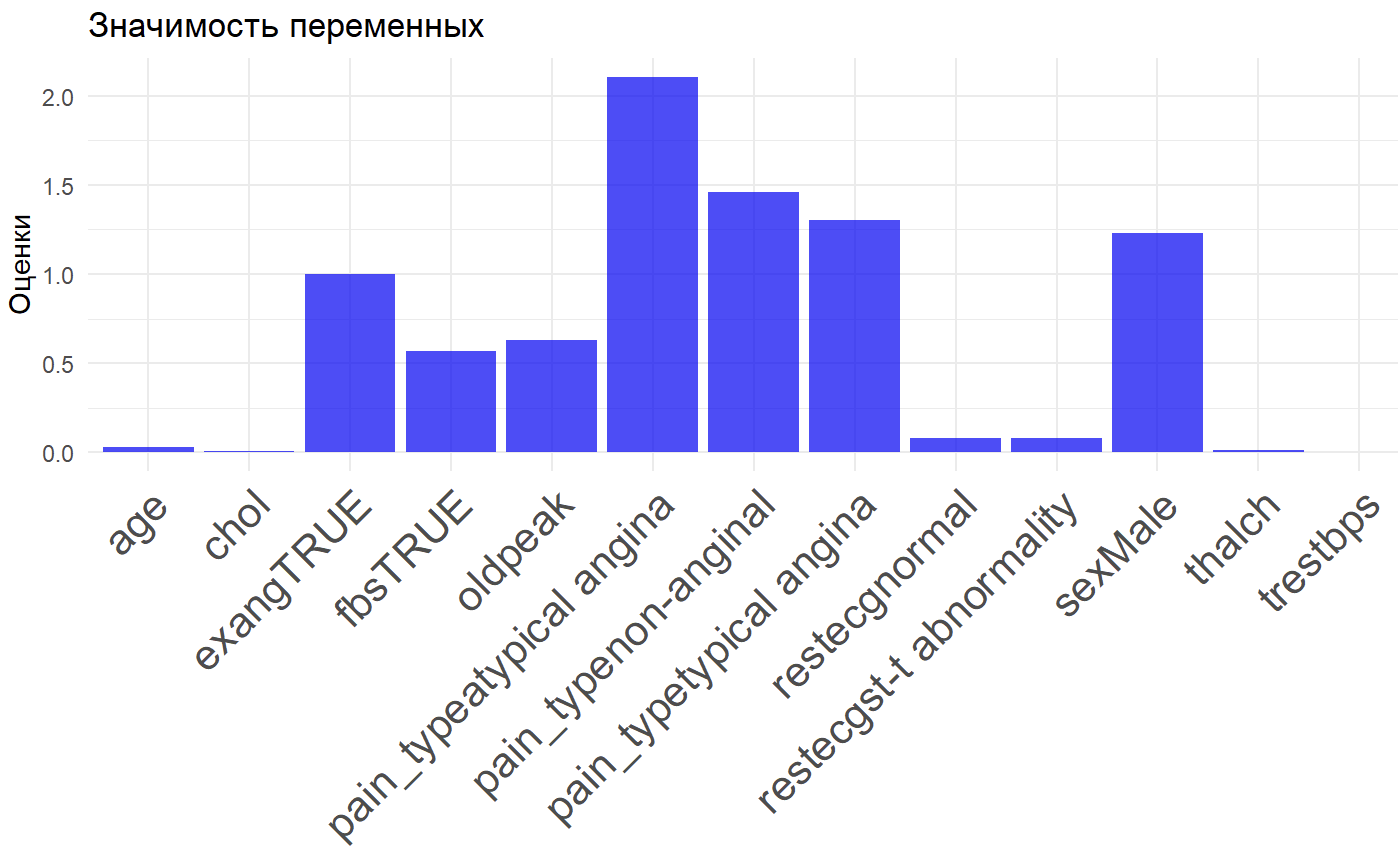


Изображение 10. При низком chol больше заболеваний

Были построены и другие графики распределений болезни с взаимодествий призаков, но никакие важные значимые взаимодествия выявлены не были

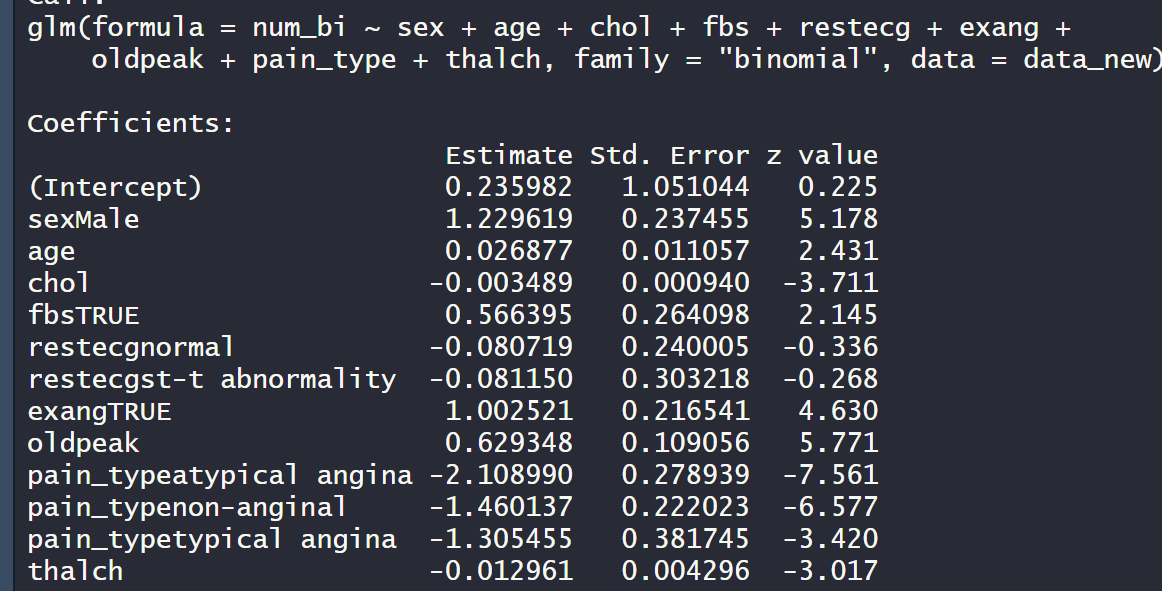


Далее была построена модель логистической регрессии для оценки риска болезни



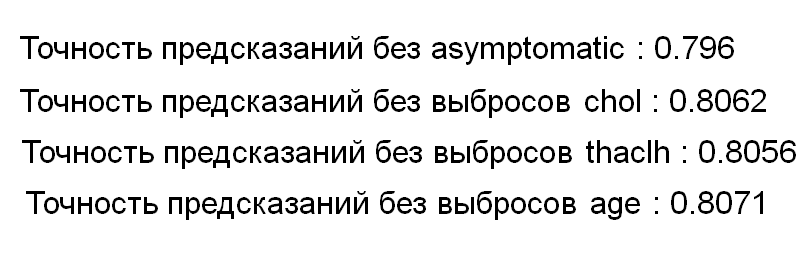
Изображение 11.График влияния значений в модели

Путем перебора признаком, было выявлено что лишь с показателем trestbps точность была ниже, что говорит о том что этим показателем можно принебречь.



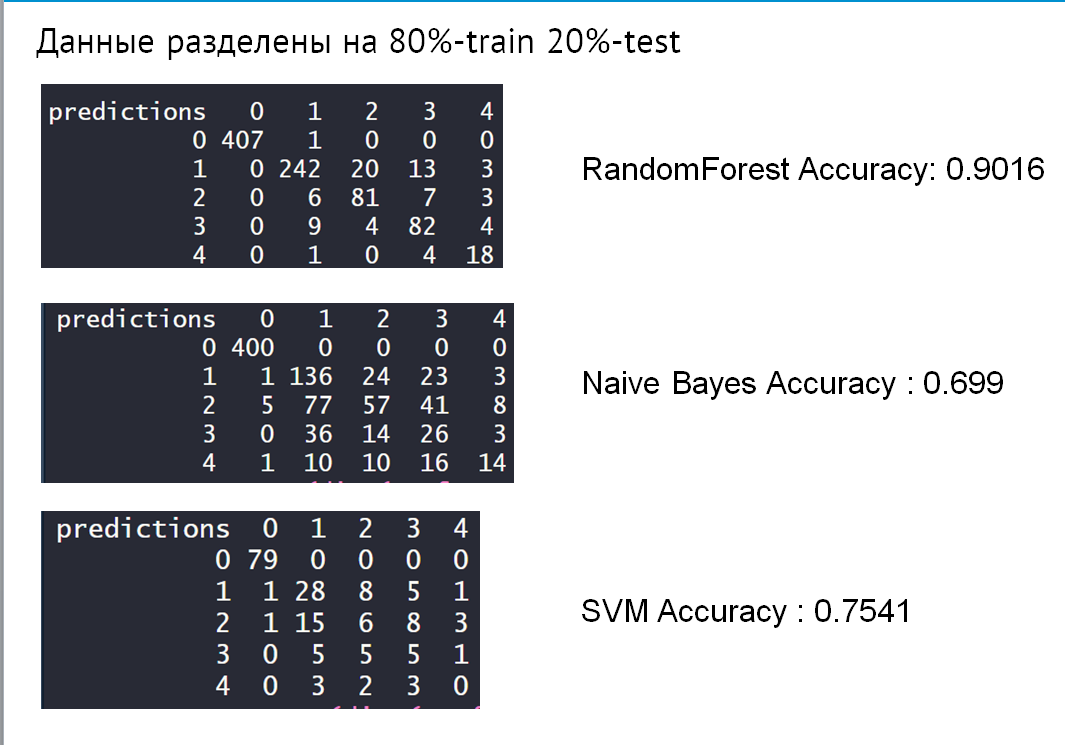
Изображение 12.Формула и показатели модели

Точность данной модели для определения риска составила 0.81

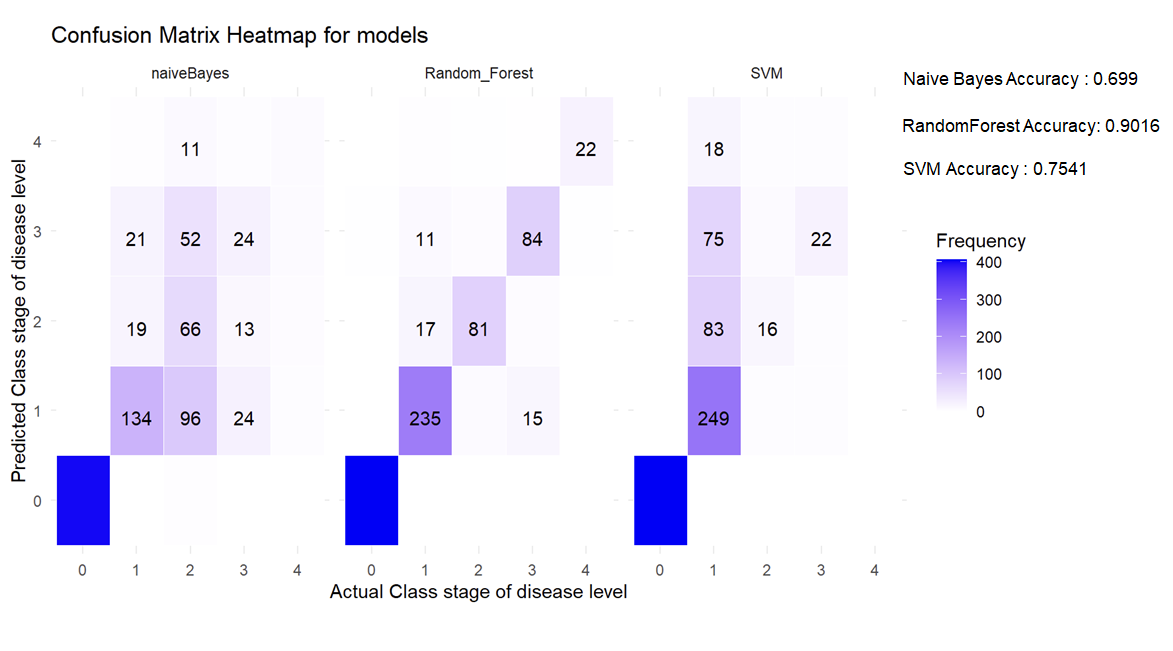


Можно сделать вывод что именно выбросы могут дать информации о наличии риска болезки, что логично, так как именно аномальные значения свидетельствуют о проблемах в организме

Так же были предприняты попытки предугадывания стадии болезни по признакам при помощи различных моделей машинного обучения, были получены следующие результаты



Изображение 13. Результаты моделей и матрицы точности



Изображение 14.Тепловая матрица для разных моделей

Для моей задачи рандомный лес деревьев решения с 5 деревьями сработал лучше всего, даже лучше чем логистическая регрессия лишь для предугадывания наличия или отстуствия болени, из чего можно сделать вывод об эффективности данного метода

GitHub - <https://github.com/Albert-Bas2002/ML-Models-Demo>

**Вывод**

Анализ датасета по сердечным заболеваниям предоставил ценные инсайты, которые имеют ключевое значение для понимания факторов риска, прогнозирования заболеваний и разработки эффективных стратегий профилактики. Полученные выводы позволяют выделить несколько важных направлений:

В первую очередь, выявление факторов риска, позволяет более точно определить основные аспекты, требующие внимания и мер по предотвращению сердечных заболеваний.

Исследование корреляций между различными параметрами открывает новые перспективы для понимания взаимодействия между различными факторами риска, что может привести к более глубоким исследованиям и новым методам профилактики.

Разработка прогностических моделей на основе данных датасета предоставляет возможность индивидуализированного подхода в медицине, что может существенно улучшить диагностику и предотвращение сердечных заболеваний.

Оценка эффективности медицинских вмешательств открывает перспективы для поиска наиболее эффективных методов лечения и профилактики, что важно для разработки оптимальных стратегий в здравоохранении.

Наконец, анализ данных о образе жизни подчеркивает важность внимания к факторам, таким как физическая активность, питание и привычки, в целях разработки всеобъемлющих программ по профилактике сердечных заболеваний.

Обобщенные выводы подчеркивают, что анализ данных по сердечным заболеваниям играет ключевую роль в научных исследованиях, обеспечивая основу для разработки эффективных подходов к управлению здоровьем. Это напоминает о важности систематического сбора и анализа данных в сфере общественного здравоохранения для создания научно обоснованных стратегий борьбы с серьезными медицинскими проблемами.

**Библиография**

1. Сердечно-сосудистые заболевания , <https://www.who.int/ru/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>

13-12-2023

1. ЗАБОЛЕВАНИЯ СЕРДЕЧНО-СОСУДИСТОЙ СИСТЕМЫ, <https://www.yamed.ru/services/terapiya/zabolevaniya-serdechno-sosudistoi-sistemy/>,

13-12-2023