Санкт-Петербургский государственный университет

Искусственный интеллект и наука о данных

Вейбер Евгения Николаевна

Финальный проект по курсу

Data Science: Инструментарий и Жизненный Цикл Проекта

Преподаватель:

Столярова Валерия Фуатовна

Санкт-Петербург

2024

Содержание работы

Введение

Целью настоящего исследования является комплексный анализ данных, полученных из известного набора данных Heart Desease (UCI), посвящённого сердечным заболеваниям [1]. Этот датасет включает в себя множество клинических параметров, которые могут быть использованы для прогнозирования наличия или отсутствия сердечного заболевания у пациентов, а также степени тяжести болезни.

Гипотеза исследования состоит в том, что определённые медицинские показатели могут коррелировать с риском развития сердечных заболеваний.

Для выполнения цели исследования были поставлены следующие задачи:

1. Оценить распределение ключевых переменных, таких как возраст, кровяное давление и холестерин, с помощью гистограмм.
2. Использовать боксплоты для визуализации различий в распределениях данных переменных между различными группами пациентов.
3. Провести анализ корреляций для количественных переменных, чтобы определить возможные связи между показателями и сердечными заболеваниями.
4. Выполнить анализ групп и кластеризацию для определения подгрупп в данных, которые могут обладать различными рисками развития заболевания.
5. Разработать модель линейной регрессии, чтобы оценить степень влияния различных факторов на вероятность возникновения сердечных заболеваний.

Ожидается, что результаты данного исследования помогут в улучшении стратегий профилактики и лечения сердечных заболеваний.

Разведочный анализ данных

Анализируя обобщенную статистику датасета, можно отметить разнообразие и распределение клинических характеристик участников. Средний возраст составляет приблизительно 53,5 лет, что указывает на выборку взрослых пациентов. При этом минимальный возраст участника исследования составил 28 лет, а максимальный — 77 лет, что свидетельствует о широком возрастном диапазоне.

Артериальное давление в покое (trestbps) и уровень холестерина в крови (chol) являются ключевыми показателями состояния сердечно-сосудистой системы. Медианные значения этих показателей находятся в пределах нормы, но максимальные значения указывают на возможное наличие гипертонической болезни и гиперлипидемии соответственно. Отсутствие данных по этим показателям (NA's) у некоторых участников требует дополнительного внимания при дальнейшем анализе.

Такие признаки, как наличие сахарного диабета (fbs), характеристики электрокардиограммы в покое (restecg), максимальная частота сердцебиений (thalch) и наличие стенокардии, вызванной упражнениями (exang), имеют бинарный характер (да/нет), что упрощает анализ влияния этих факторов на исследуемые исходы.

Симптомы, такие как депрессия ST-сегмента (oldpeak) и наклон ST-сегмента во время упражнений (slope), варьируются у участников и могут указывать на различные степени риска ишемической болезни сердца.

В исследовании также рассматривается количество основных коронарных артерий, заблокированных кальцием (ca), и тип метаболического агента (thal), который также представлен в виде категориальных данных, но имеет большое количество отсутствующих значений, что ограничивает их использование в некоторых анализах.

Целевой переменной в датасете является num, представляющая степень сердечного заболевания от 0 (отсутствие заболевания) до 4

(максимальное выраженное заболевание). Среднее значение этого признака близко к 1, что указывает на наличие сердечных заболеваний у значительной части участников исследования.

В целом, представленные данные дают основу для детального исследования факторов риска развития сердечно-сосудистых заболеваний и могут использоваться для создания моделей прогнозирования здоровья сердца. Однако для точных выводов необходимо дальнейшее статистическое анализирование и возможно, очистка данных от пропусков и выбросов.

После заполнения отсутствующих значений в данных, удаления малоинформативных колонок id, dataset, ca и конвертации категориальных переменных в факторы, набор данных приобрел следующий вид (рис. 1) с 920 записями и 13 переменными.

```
> head(data)
# A tibble: 6 × 13
    age sex    cp              trestbps  chol fbs   restecg        thalch exang oldpeak slope       thal             num
  <dbl> <fct>  <fct>              <dbl> <dbl> <fct> <fct>           <dbl> <fct>   <dbl> <fct>       <fct>          <dbl>
1    63 Male   typical angina       145   233 TRUE  lv hypertrophy    150 FALSE     2.3 downsloping fixed defect       0
2    67 Male   asymptomatic         160   286 FALSE lv hypertrophy    108 TRUE      1.5 flat        normal             2
3    67 Male   asymptomatic         120   229 FALSE lv hypertrophy    129 TRUE      2.6 flat        reversable de…     1
4    37 Male   non-anginal          130   250 FALSE normal            187 FALSE     3.5 downsloping normal             0
5    41 Female atypical angina      130   204 FALSE lv hypertrophy    172 FALSE     1.4 upsloping   normal             0
6    56 Male   atypical angina      120   236 FALSE normal            178 FALSE     0.8 upsloping   normal             0
```

Рисунок 1. Данные после подготовки

Гистограмма распределения возраста (age) показывает, что наличие и тяжесть сердечных заболеваний коррелирует с более старшим возрастом (рис. 2).

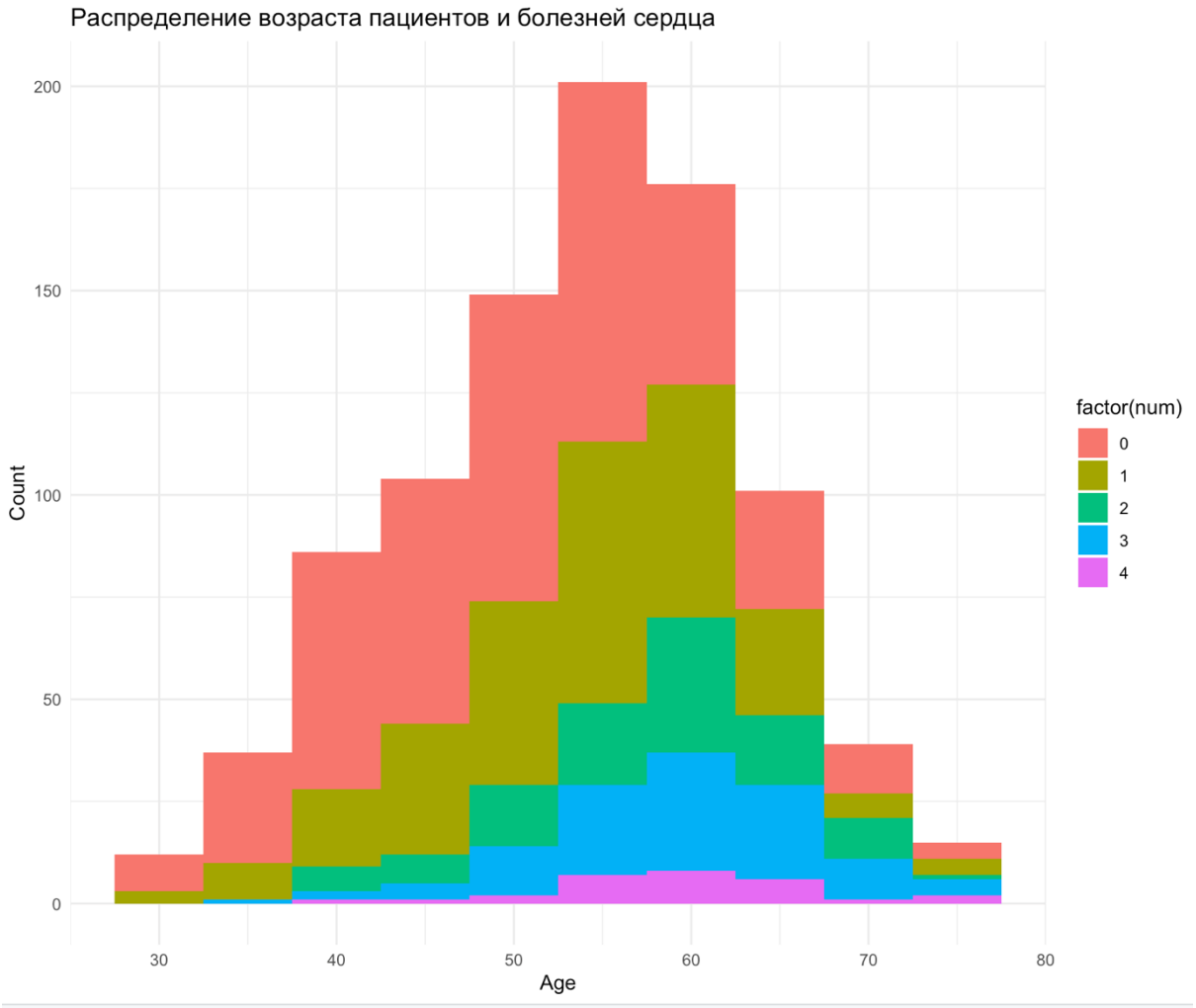

Рисунок 2. Распределение возраста и болезней сердца

Столбчатая диаграмма распределения типов болей в груди (cp) показывает, что чаще всего при заболеваниях сердца боль в груди имеет тип “asymptomatic” (бессимптомная). Однако, при отсутствии сердечных заболеваний часто могут наблюдаться такие типы болей в груди, как “atypical angina” (атипичная стенокардия), “non-anginal” (неангинальная) (рис. 3).

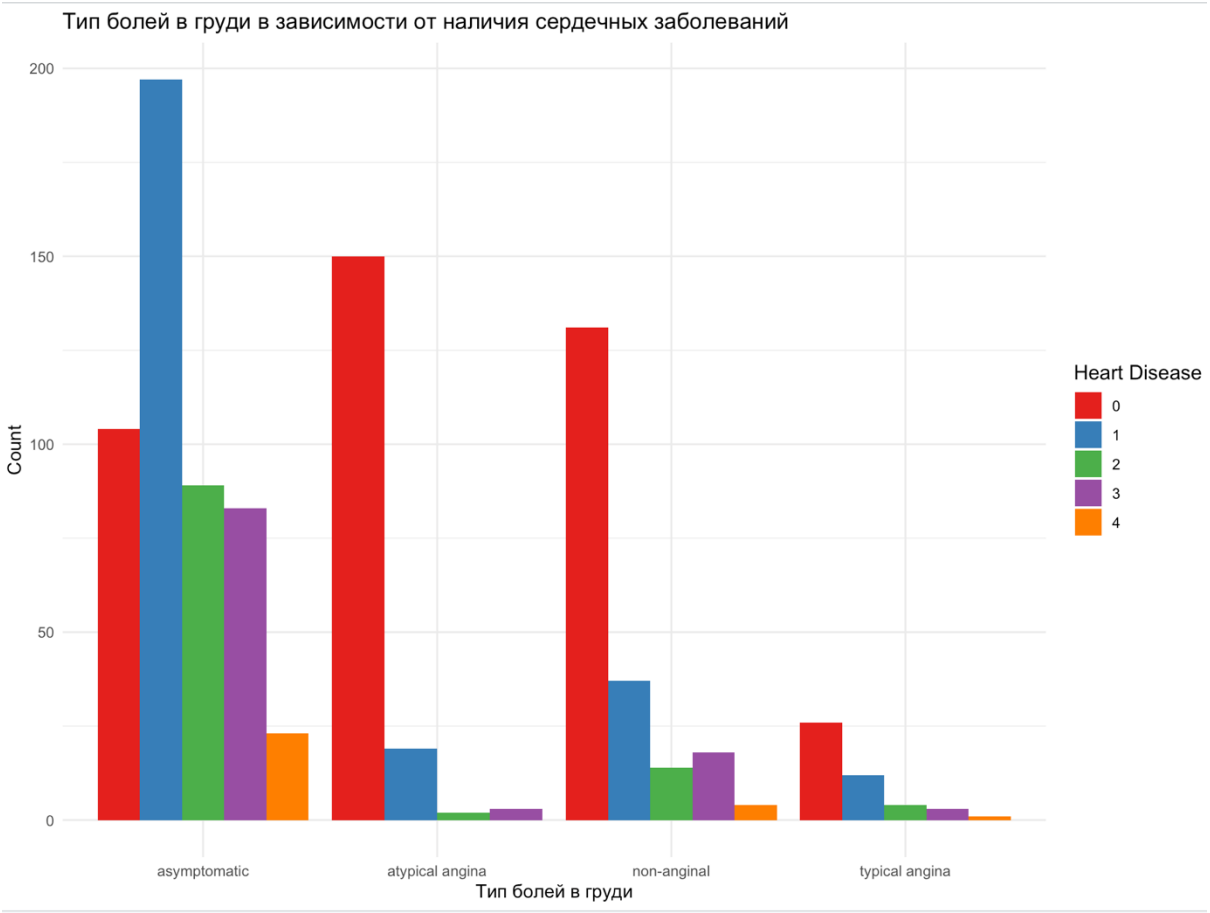

Рисунок 3. Тип болей в груди при сердечных заболеваниях

Столбчатая диаграмма распределения полов (sex) показывает, что мужчины больше, чем женщины, подвержены заболеваниям сердца (рис. 4).

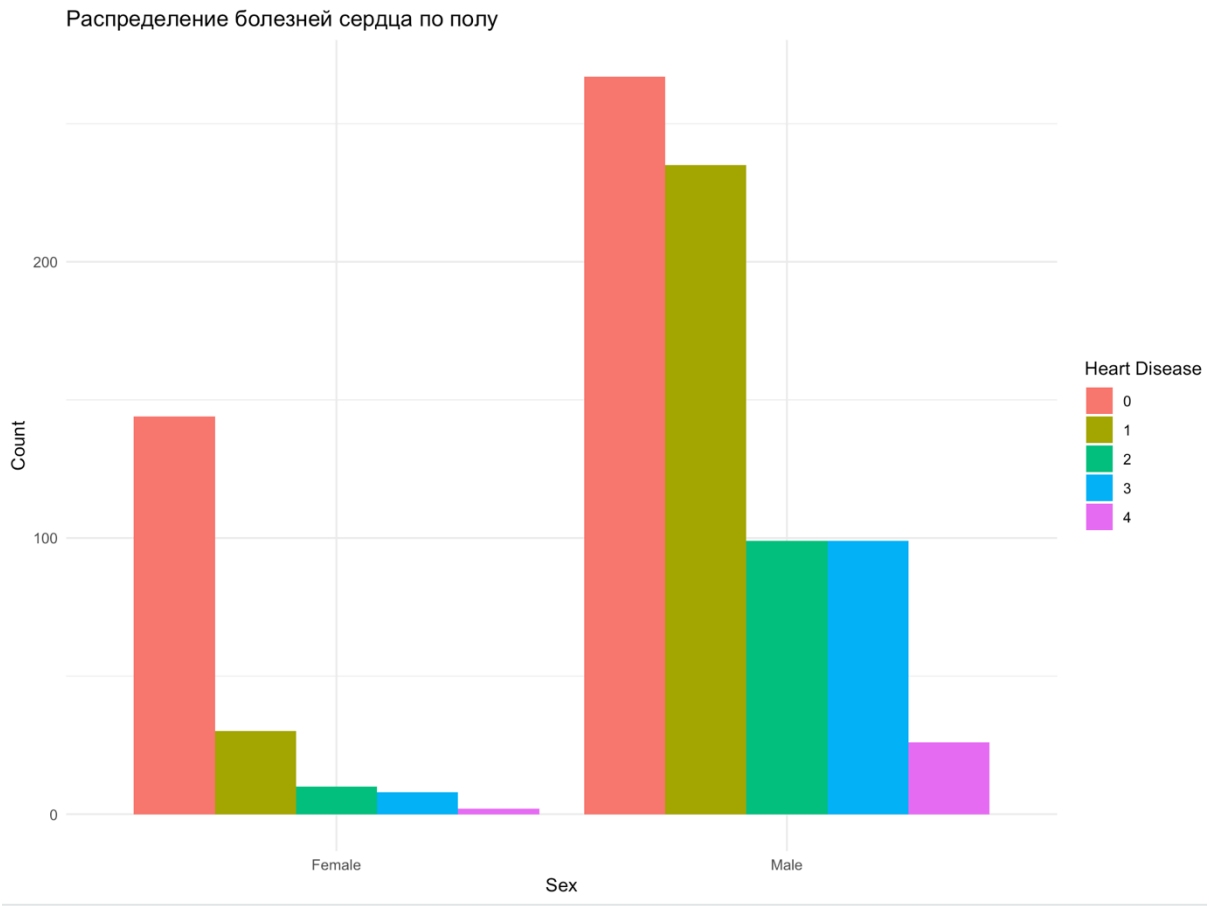

Рисунок 4. Распределение болезней сердца по полу

Столбчатая диаграмма распределения уровня сахара в крови натощак > 120mg/dl (fbs) показывает, что высокий уровень этого показателя не коррелирует с заболеваниями сердца (рис. 5).

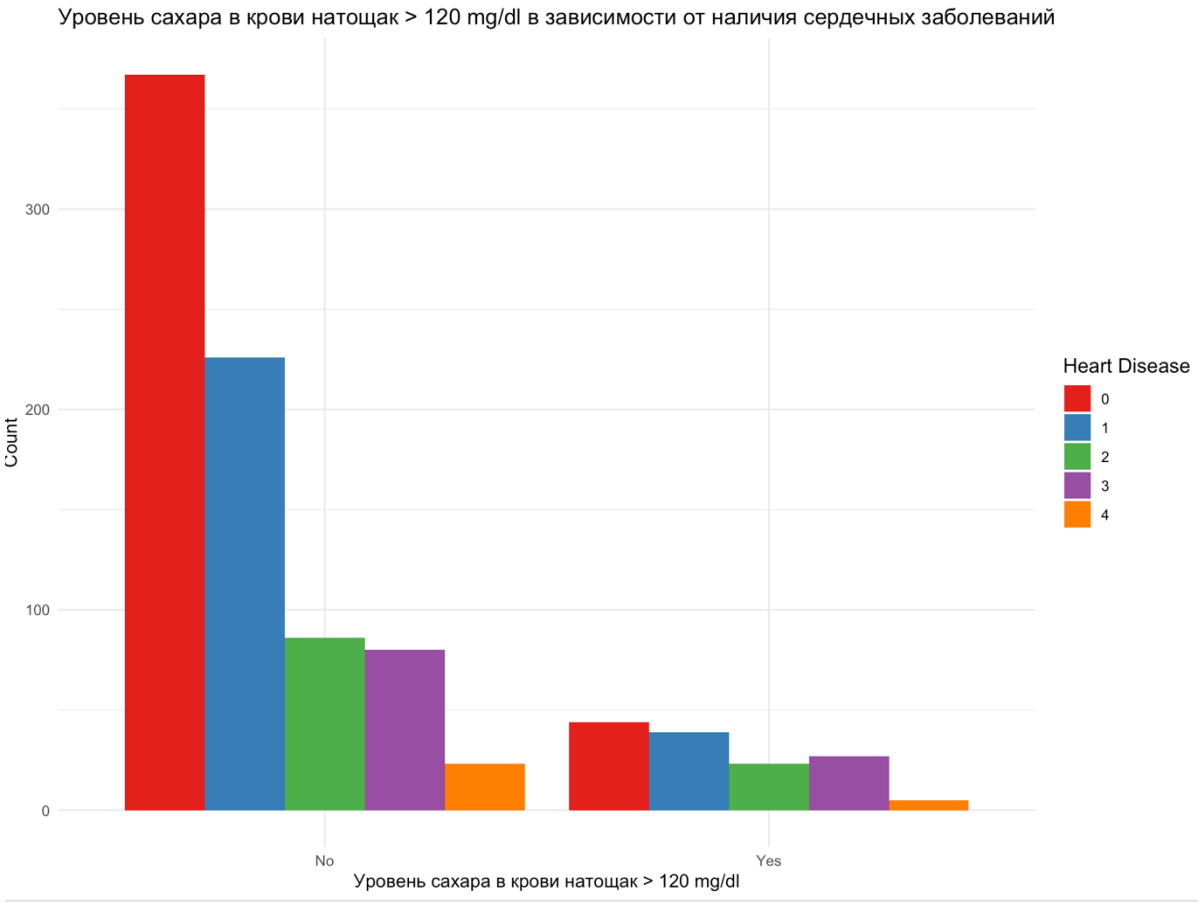

Рисунок 5. Уровень сахара>120mg/dl и заболевания сердца

Столбчатая диаграмма наличия стенокардии, вызванной физической нагрузкой (exang) показывает, что при заболеваниях сердца корреляции с этим показателем не наблюдается, однако у пациентов без заболеваний сердца данного симптома нет в большинстве случаев (рис. 6).

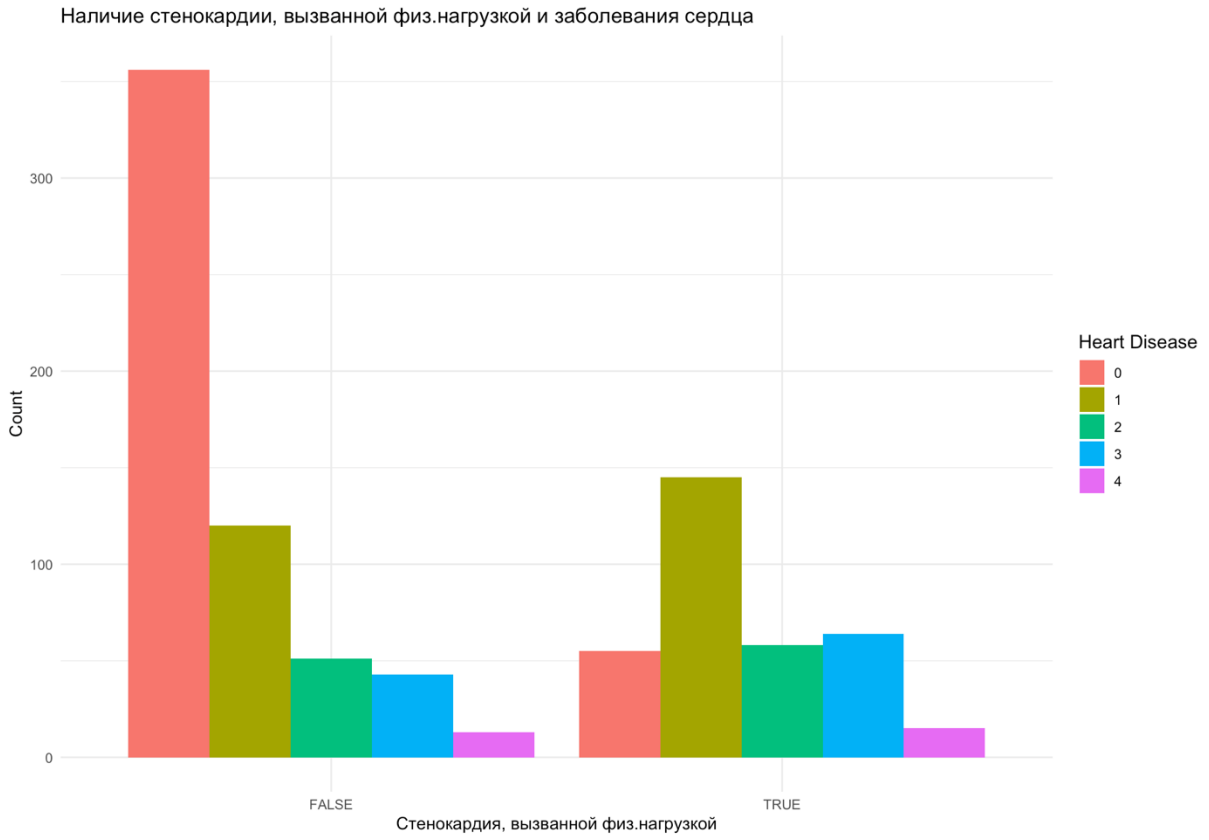

Рисунок 6. Наличие стенокардии, вызванной физ. нагрузкой, и заболевания сердца

График «ящик с усами» распределения уровня холестерина (chol) при заболеваниях сердца не показал наличия корреляций (рис. 7).

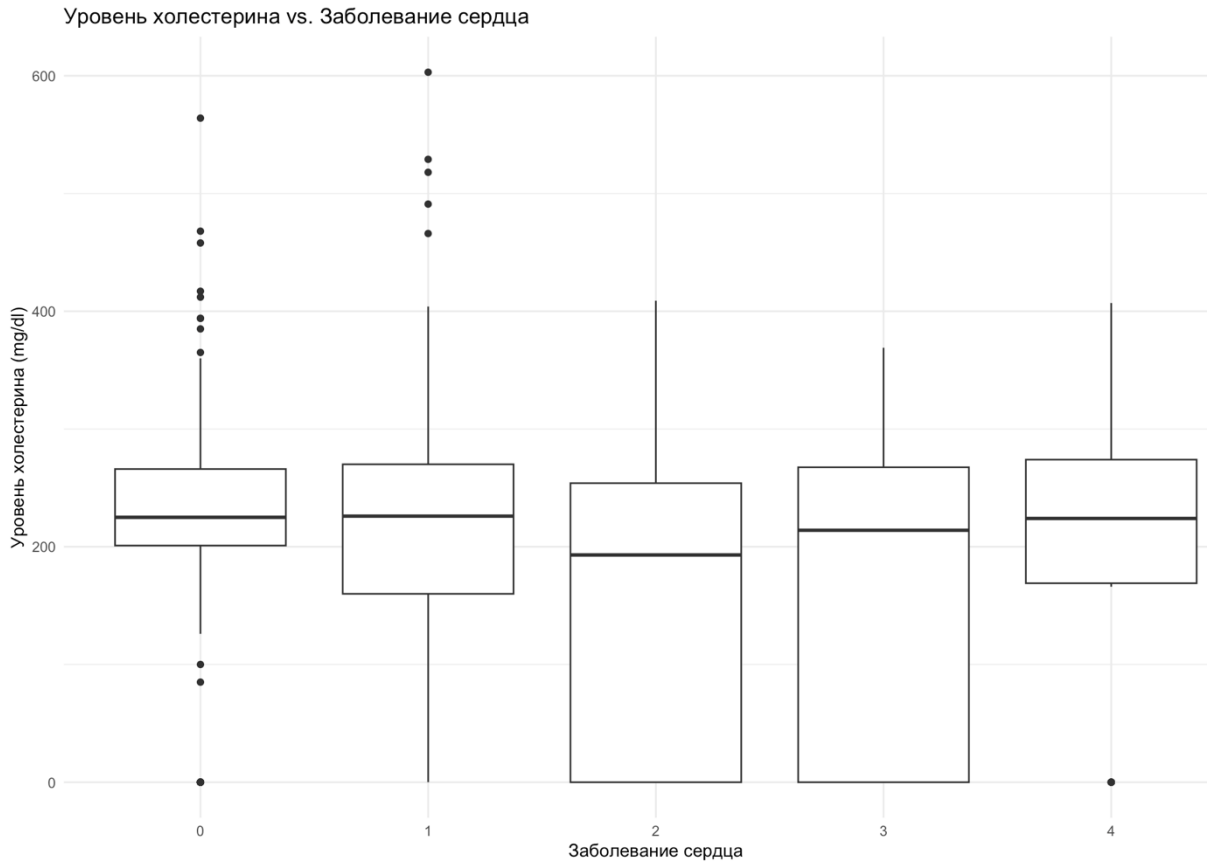

Рисунок 7. Уровень холестерина и заболевания сердца

Точечная диаграмма распределения максимальной частоты сердечного ритма (thalch) с возрастом (age) при заболеваниях сердца показала снижение максимальной частоты сердечного ритма и приобретение/усугубление сердечных заболеваний с возрастом (рис. 8).

Рисунок 8. Максимальная частота сердечного ритма с возрастом и заболевания сердца

Столбчатая диаграмма распределения среднего значения артериального давления в покое (trestbps) при заболеваниях сердца не показала наличия корреляций (рис. 9).

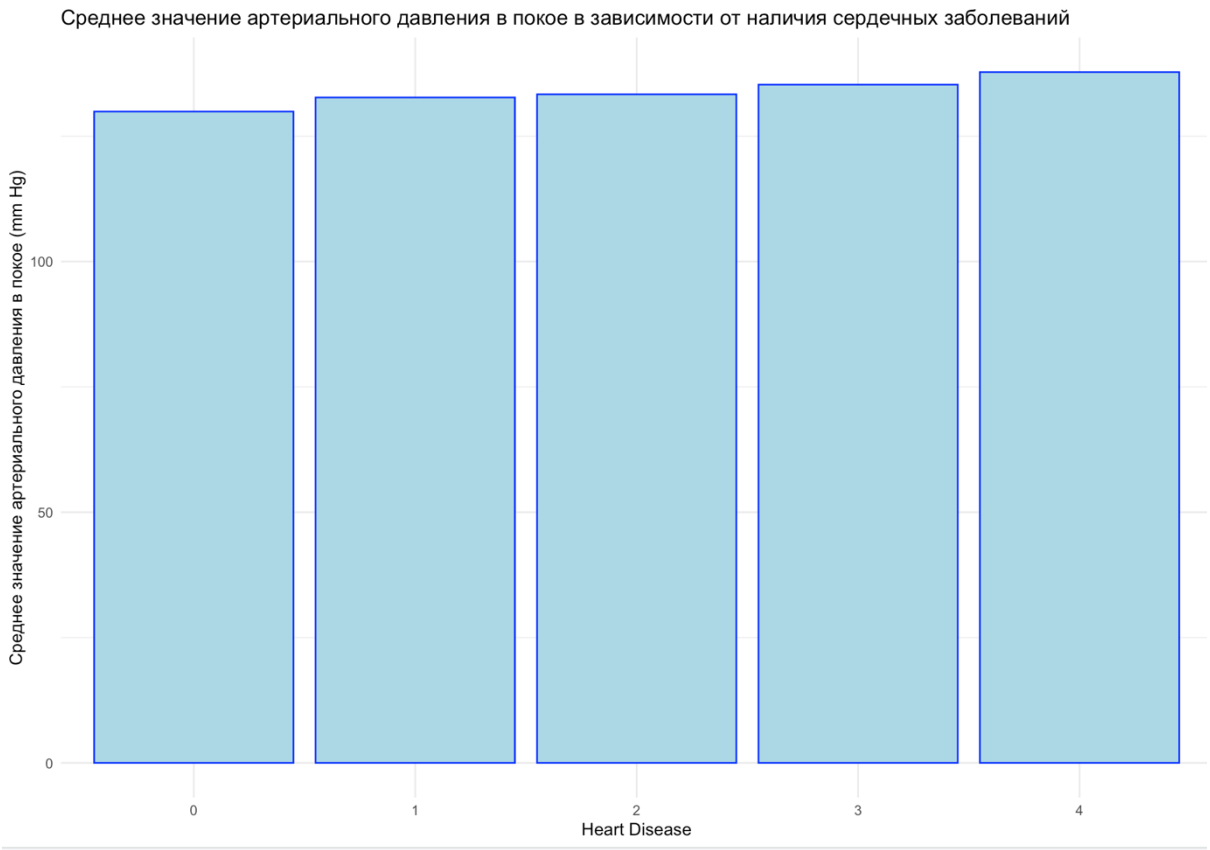

Рисунок 9. Среднее значение артериального давления в покое

Столбчатая диаграмма распределения результатов электрокардиографии в покое (restesg) при заболеваниях сердца не показала предсказательного потенциала при отклонениях в этом показателе (рис. 9).

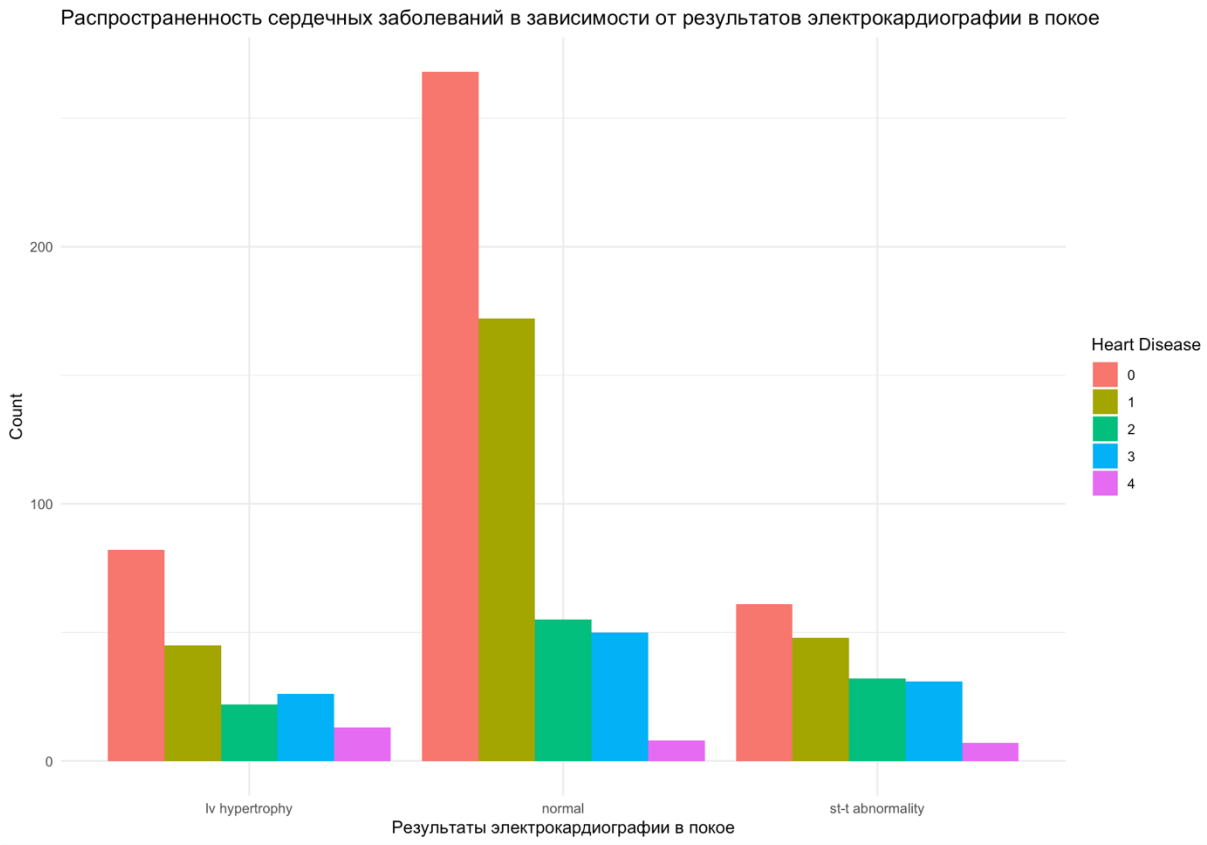

Рисунок 10. Результаты электрокардиографии в покое при заболеваниях сердца

Анализ корреляций

Для выполнения анализа корреляций категориальные переменные были преобразованы в численные.

Получившийся график корреляций (рис. 11) и корреляционная матрица (рис. 12) свидетельствуют о том, что между целевым признаком num и независимыми признаками существуют следующие корреляции:

1) Умеренная положительная корреляция с показателем возраста age (0.34) указывает, что с увеличением возраста увеличивается и значение num, что может свидетельствовать о повышении риска сердечных заболеваний с возрастом.
2) Умеренная отрицательная корреляция с показателем уровня холестерина chol (-0.23) может указывать на то, что более низкие уровни холестерина ассоциируются с более серьезными уровнями сердечных заболеваний num, что противоречит общепринятым представлениям о роли холестерина в развитии сердечных заболеваний.
3) Умеренная отрицательная корреляция с переменной thalch (-0.35) может свидетельствовать о том, что более высокий максимальный уровень сердцебиения при нагрузке связан с более низким риском сердечных заболеваний.
4) Умеренная положительная корреляция с переменной oldpeak (0.41) указывает на более высокий риск сердечных заболеваний при большей степени депрессии ST-сегмента, вызванной упражнениями относительно отдыха.
5) Умеренная положительная корреляция с показателем пола sex (0.26) указывает, что для мужчин увеличивается риск приобретения и усугубления сердечных заболеваний num.

6) Умеренная отрицательная корреляция с переменной cp (-0.35) может свидетельствовать о том, что при заболеваниях сердца чаще встречается бессимптомная и типичная стенокардия.
7) Умеренная положительная корреляция с показателем наличия стенокардии, вызванной физическими упражнениями exang (0.35) указывает на то, что при наличии стенокардии, вызванной упражнениями, повышается вероятность наличия заболевания сердца.

```
> print(cor_matrix)
                          age     trestbps        chol      thalch     oldpeak         num
age             1.000000000  0.230783971 -0.08600982 -0.34971486  0.23355008  0.33959559
trestbps        0.230783971  1.000000000  0.08948440 -0.10474715  0.16121750  0.11317825
chol           -0.086009819  0.089484396  1.00000000  0.22604734  0.04745372 -0.23053946
thalch         -0.349714865 -0.104747152  0.22604734  1.00000000 -0.14940057 -0.34917315
oldpeak         0.233550076  0.161217497  0.04745372 -0.14940057  1.00000000  0.41158800
num             0.339595593  0.113178253 -0.23053946 -0.34917315  0.41158800  1.00000000
sex_numeric     0.056889351 -0.002144745 -0.19402890 -0.17459004  0.09079398  0.25934161
cp_numeric     -0.214352435 -0.072079031  0.12871485  0.26743163 -0.29156319 -0.37738226
fbs_numeric     0.219914829  0.143709319  0.08931633 -0.02647210  0.03411190  0.12953878
restecg_numeric -0.006651533  0.010262957 -0.19959379 -0.16181798 -0.04723488  0.03460349
exang_numeric   0.160910345  0.153835278 -0.03391372 -0.35417280  0.39564953  0.35056653
slope_num      -0.077086601 -0.050445905  0.03322085  0.28925097 -0.25622268 -0.20187484
thal_num        0.101109443  0.066675247 -0.05270751 -0.05433643  0.12214662  0.18638691
```

Рисунок 11. Матрица корреляций

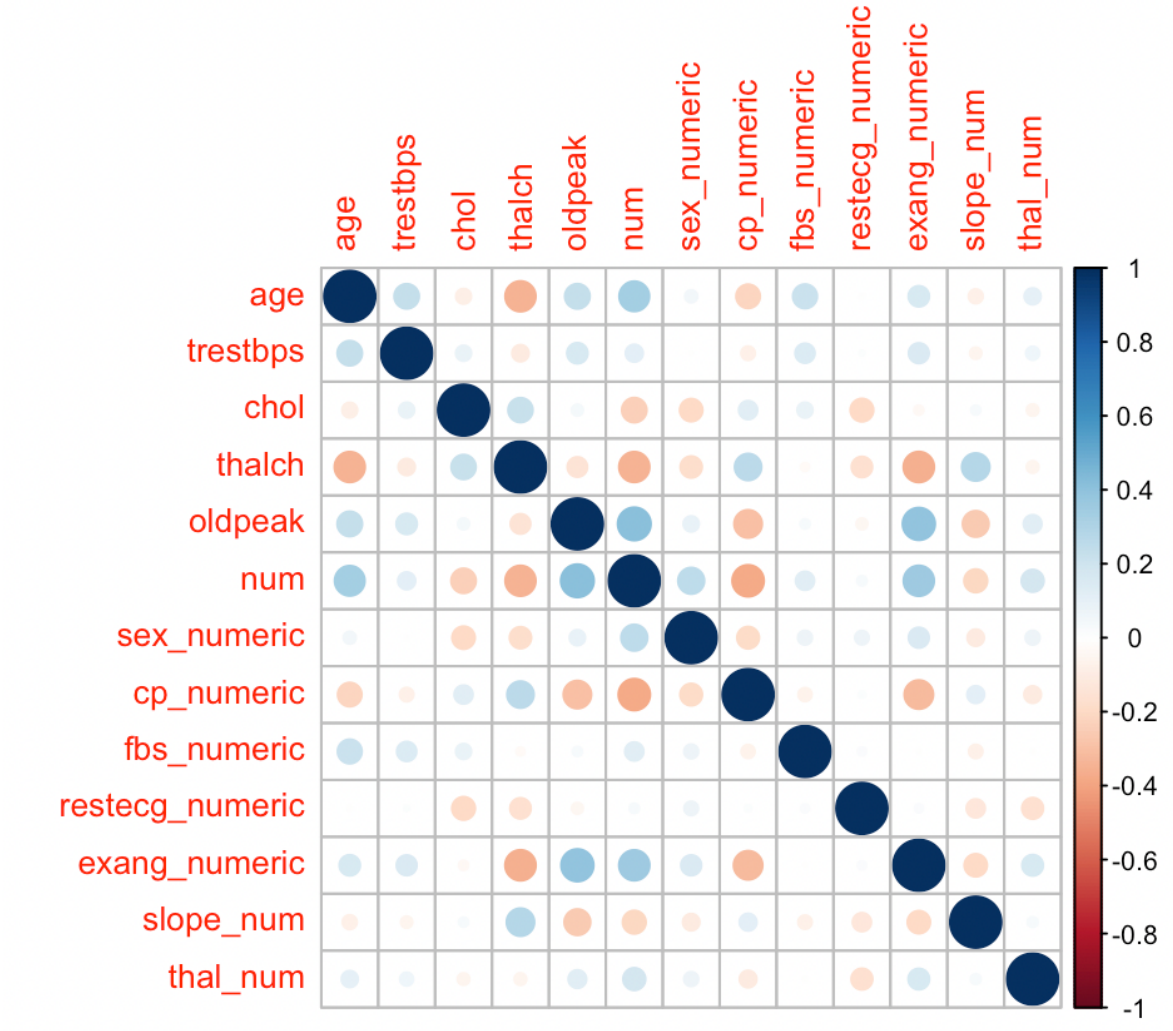

Рисунок 12. График корреляций

Таким образом, было отмечено, что заметные положительные коэффициенты корреляции с зависимым признаком имеют переменные age (возраст), oldpeak (депрессия ST-сегмента, вызванная физической нагрузкой), sex (пол), exang (наличие стенокардии). Заметные отрицательные коэффициенты корреляции имеют переменные chol (уровень холестерина), cp (тип болей в груди).

Статистический вывод

Для независимых переменных, показавших значительные коэффициенты корреляции с зависимой переменной num был проведен статистический вывод с целью установить, являются ли данные корреляции значимыми.

1) Возраст (age) и степень заболевания сердца (num):
   1.1) Коэффициент корреляции Пирсона равен примерно 0.34, что указывает на умеренную положительную линейную взаимосвязь между возрастом и степенью сердечного заболевания. Это означает, что с увеличением возраста склонность к более высокой степени сердечного заболевания также увеличивается.
   1.2) t-значение равно 10.939, что указывает на сильную статистическую значимость наблюдаемой корреляции.
   1.3) P-значение меньше 2.2e-16, что значительно меньше общепринятого порога значимости 0.05. Это означает, что результаты являются статистически значимыми, и вероятность получить такую сильную корреляцию случайно крайне мала.
   1.4) 95% доверительный интервал коэффициента корреляции находится между 0.281 и 0.396, что означает, что мы можем быть на 95% уверены, что в общей популяции значение коэффициента корреляции Пирсона между age и num будет в пределах этого интервала.

   Таким образом, эти результаты подтверждают наличие статистически значимой положительной связи между возрастом и степенью сердечного заболевания в исследуемой выборке.
2) Уровень холестерина (chol) и степень заболевания сердца (num):

2.1) Коэффициент корреляции Пирсона равен примерно -0.23, что указывает на слабую отрицательную линейную взаимосвязь между уровнем холестерина и степенью сердечного заболевания. Это означает, что с увеличением уровня холестерина наблюдается тенденция к более низкой степени сердечного заболевания (или наоборот, более высокие значения `num` связаны со сниженным уровнем холестерина).

2.2) t-значение равно -7.1784, что является высоким по абсолютному значению и указывает на сильную статистическую значимость обнаруженной корреляции.

2.3) P-значение равно 1.455e-12, что значительно меньше стандартного порога в 0.05, свидетельствуя о статистической значимости результатов.

2.4) 95% доверительный интервал для коэффициента корреляции находится между -0.291 и -0.168, что подтверждает, что в общей популяции значение коэффициента корреляции будет в пределах этого интервала с 95% вероятностью.

Таким образом, результаты анализа показывают статистически значимую обратную связь между уровнем холестерина и степенью сердечного заболевания в исследуемой выборке. Это может указывать на то, что пациенты с более высоким уровнем холестерина имеют меньшую вероятность тяжелых форм сердечных заболеваний, хотя для более точных выводов необходимо учитывать и другие факторы, а также провести дополнительный анализ причинно-следственных связей.

3) Максимальный пульс, достигнутый во время нагрузки (thalch) и степень заболевания сердца (num):

3.1) Коэффициент корреляции Пирсона приблизительно равен -0.349, что указывает на умеренную обратную линейную

взаимосвязь между максимальными пульсом и степенью сердечного заболевания. С увеличением пульса наблюдается тенденция к более низкой степени сердечных заболеваний, или, наоборот, более серьёзные степени сердечных заболеваний связаны с более низким максимальным пульсом.

3.2) Значение t-статистики составляет -11.29, что говорит о значительном отклонении от нуля и свидетельствует о статистической значимости найденной корреляции.

3.3) P-значение меньше 2.2e-16, что значительно меньше принятого порога в 0.05, подтверждая статистическую значимость результатов.

3.4) 95% доверительный интервал для коэффициента корреляции находится в диапазоне от -0.405 до -0.291, что уверенно подтверждает наличие обратной связи в генеральной совокупности с 95% вероятностью.

Эти результаты могут свидетельствовать о том, что пациенты с более низким уровнем максимального пульса во время нагрузки имеют более высокий риск сердечных заболеваний. Это может быть связано с пониженной способностью сердца справляться с физической нагрузкой, что часто встречается при различных сердечно-сосудистых расстройствах.

4) Депрессия ST-сегмента, вызванная упражнениями относительно состояния покоя (oldpeak) и уровень сердечного заболевания (num):

4.1) Коэффициент корреляции Пирсона приблизительно равен 0.412. Это означает, что существует средняя положительная линейная связь между величиной депрессии ST-сегмента и степенью сердечного заболевания. Другими словами, более высокие значения

oldpeak ассоциируются с более серьезными сердечными заболеваниями.

4.2) Высокое значение t-статистики (приблизительно 13.683) и очень низкое p-значение (меньше 2.2e-16) надежно свидетельствуют о статистической значимости этой взаимосвязи.

4.3) Доверительный интервал коэффициента корреляции, который не включает ноль и находится в диапазоне от 0.356 до 0.464, дополнительно подтверждает уверенность в наличии положительной корреляции.

На основе этих результатов можно заключить, что изменение величины депрессии ST-сегмента статистически значимо связано с изменением риска сердечных заболеваний.

5) Пол (sex) и уровень сердечного заболевания (num):

5.1) Коэффициент корреляции Пирсона приблизительно равен 0.259. Это указывает на то, что пол пациента может быть связан с риском развития сердечного заболевания, хотя эта связь не является сильной.

5.2) Значение t-статистики равное 8.136 и p-значение, которое гораздо меньше стандартного порога 0.05, свидетельствуют о статистической значимости этой корреляции.

5.3) Доверительный интервал коэффициента корреляции, не включающий ноль и находящийся между 0.198 и 0.319, подкрепляет вывод о статистической уверенности в наличии данной корреляции.

Таким образом, можно предположить, что половая принадлежность имеет влияние на вероятность развития

сердечного заболевания, и этот фактор можно учитывать в комплексной оценке сердечно-сосудистых рисков.

6) Тип болей в груди (cp) и сердечные заболевания (num):
   6.1) Корреляционный анализ Пирсона для переменных cp_numeric и num выявил отрицательную связь с коэффициентом корреляции примерно -0.377. Основываясь на предыдущих операциях преобразования качественной переменной cp в количественный формат, можно сделать вывод о том, что для значений cp = “typical angina” (типичная стенокардия) и cp = “asymptomatic” (бессимптомная) вероятность заболеваний сердца выше, чем при других значениях cp.
   6.2) Значение t-статистики равное -12.347 и p-значение, стремящееся к нулю, свидетельствуют о том, что эта обратная корреляция статистически значима.
   6.3) Доверительный интервал для коэффициента корреляции от -0.431 до -0.321 не включает ноль, что усиливает доверие к тому, что обнаруженная связь не является случайной.

   Таким образом, корреляцию между типичной/бессимтомной стенокардией и повышенным риском болезней сердца можно считать статистически значимой.

7) Стенокардия, вызванная упражнениями (exang) и болезни сердца (num):
   7.1) Анализ корреляции Пирсона показывает положительный коэффициент корреляции в размере примерно 0.350. Это означает, что между наличием стенокардии при упражнениях и уровнем сердечного заболевания существует умеренная положительная связь.

7.2) Значение t-статистики равно 11.341, что является высоким значением и указывает на значительную статистическую разницу от нуля.

7.3) P-значение меньше 2.2e-16, что гораздо ниже общепринятого порога значимости 0.05, подтверждает, что результат не случаен.

7.4) Доверительный интервал коэффициента корреляции от 0.293 до 0.406 также не включает ноль и указывает на уверенность в наличии этой положительной связи.

Эти результаты предполагают, что у пациентов, испытывающих стенокардию при физических нагрузках, вероятно, выше уровень сердечных заболеваний, что может быть важным индикатором при оценке состояния сердечно-сосудистой системы.

Подводя итог, нужно отметить, что переменные, показавшие сильную или умеренную корреляцию с зависимой переменной, оказывают на нее статистически значимый эффект. В число этих переменных вошли age, chol, thalch, oldpeak, sex, cp, exang.

Анализ групп

Для проведения анализа групп данные были разделены на 5 групп в соответствии со значениями целевой переменной num (0, 1, 2, 3, 4).

Группа num = 0 в наборе данных представляет пациентов без сердечных заболеваний. Распределение возраста здесь варьируется от 28 до 76 лет, с медианным возрастом 51 год. Это показывает, что люди всех возрастных групп представлены в этой категории, но большинство пациентов среднего возраста. В группе 267 мужчин и 144 женщины, что указывает на большее количество мужчин. Большинство пациентов описывают боли (cp) как атипичную стенокардию (150) или неангинальную боль (131), что может указывать на меньшую вероятность серьезных сердечных проблем в этой группе. Артериальное давление (trestbps) варьируется от 80 до 190 мм рт. ст., с медианным значением 130 мм рт. ст., что в пределах нормы. Уровень холестерина (chol) имеет очень широкий диапазон от 0 до 564 мг/дл, с медианой 225 мг/дл. Некоторые очень высокие значения могут требовать дополнительного анализа на предмет ошибок или особых случаев. Большинство (367) пациентов имеют нормальный уровень сахара в крови натощак (fbs), в то время как 44 имеют повышенный уровень. Максимальный пульс при нагрузке (thalch) имеет распределение от 69 до 202 уд/мин, с медианой 150 уд/мин, что является хорошим показателем способности к физической нагрузке. У большинства (356) нет стенокардии при нагрузке (exang), что хорошо коррелирует с отсутствием сердечных заболеваний. Депрессия ST-сегмента (oldpeak) в данной группе показывает незначительные изменения от -1.1 до 4.2, что указывает на различные уровни сердечной нагрузки, но в целом малозначительные для данной группы. Большинство имеет нормальную ЭКГ (restecg) (268), в то время как другие показывают гипертрофию левого желудочка (82) или нарушения ST-T (61).

Группа с num = 0 показывает отсутствие клинически значимых сердечных заболеваний среди исследуемых. Результаты подчеркивают важность учета как демографических, так и медицинских переменных при оценке сердечного здоровья. Возможно, высокие значения холестерина и расширенный диапазон артериального давления заслуживают дополнительного внимания даже среди пациентов без очевидных сердечных заболеваний.

Группа num = 1 в наборе данных представляет пациентов с установленными сердечными заболеваниями первой степени тяжести. Возрастные показатели (age) в ней варьируются от 31 до 75 лет, с медианным возрастом 55 лет. Это подчеркивает, что сердечные заболевания чаще встречаются у пациентов среднего и старшего возраста. Группа включает значительно больше мужчин (235) по сравнению с женщинами (30), что может указывать на более высокую распространенность сердечных заболеваний среди мужчин в данной выборке. Большинство пациентов описывают свои боли (cp) как бессимптомные (197), что указывает на более серьезное состояние, связанное с сердечными заболеваниями. Наблюдается широкий диапазон показаний артериального давления (trestbps): от 92 до 200 мм. рт.ст., с медианным значением 130 мм. рт.ст. Высокие максимальные значения могут указывать на гипертонические состояния среди этих пациентов. Значения холестерина (chol) колеблются от 0 до 603 мг/дл, с медианным значением 226 мг/дл. Значения максимальной ЧСС (thalch) от 72 до 195 уд/мин, с медианой 130 уд/мин. Низкие максимальные значения могут указывать на ограниченную способность сердца к работе под нагрузкой. У большинства пациентов (145 из 265) есть стенокардия, вызванная упражнениями (exang), что является серьезным индикатором наличия ишемической болезни сердца. Депрессия ST-сегмента (oldpeak) принимает значения от -2.6 до 5.0, что указывает на значительные изменения, связанные с ишемией сердечной мышцы при

нагрузках. Большинство имеют нормальную ЭКГ (restecg) (172), хотя у значительного числа пациентов (93) присутствуют изменения, связанные с гипертрофией левого желудочка или аномалиями ST-T.

Группа с num = 1 демонстрирует различные клинические признаки и показатели, типичные для пациентов с сердечными заболеваниями, включая высокий уровень стенокардии при нагрузках, изменения на ЭКГ и значительные колебания артериального давления и холестерина. Эти данные могут быть использованы для дальнейшего изучения связи между клиническими признаками и степенью сердечных заболеваний.

Группа num = 2 представляет пациентов с умеренной степенью сердечного заболевания. Анализируя представленные характеристики этой группы, можно сделать следующие наблюдения. Пациенты в этой группе находятся в возрастном диапазоне (age) от 38 до 74 лет, с медианой в 58 лет. Большинство пациентов приближаются к пожилому возрасту, что соответствует общему риску увеличения сердечных заболеваний с возрастом. В группе значительно больше мужчин (99) по сравнению с женщинами (10), что может отражать более высокий риск сердечных заболеваний среди мужчин в данной выборке. Большинство пациентов испытывают бессимптомную боль в груди (89 из 109) (cp), что указывает на серьёзность ишемических проявлений у этих пациентов. Артериальное давление (trestbps) колеблется от 95 до 180 мм рт.ст., с медианой в 130 мм рт.ст., что входит в пределы нормы для взрослого населения. Наблюдается очень широкий диапазон уровней холестерина (chol) от 0 до 409 мг/дл. Значения максимального пульса (thalch) от 60 до 180 уд/мин, с медианой 130 уд/мин, что может указывать на ограниченную способность сердца к работе под нагрузкой у некоторых пациентов. Стенокардия при нагрузке (exang) наблюдается у 58 пациентов, что может быть индикатором наличия более серьезных сердечных заболеваний. Значения депресии ST-сегмента (oldpeak) от -2 до 4, с медианой в 1.4, указывают на значительные изменения

при нагрузочных тестах, что коррелирует с наличием сердечных заболеваний. Наличие гипертрофии левого желудочка (22), нормальные показатели (55) и нарушения ST-T (32) отражают различные сердечные состояния в этой группе по результатам электрокардиограммы (restecg).

Группа с num = 2 включает пациентов с умеренной степенью сердечного заболевания, что проявляется в серьезных симптомах и клинических показателях. Большинство пациентов испытывают бессимптомную или неангинальную боль, что требует дальнейшего внимания к диагностике и лечению.

Группа num = 3 включает пациентов с тяжелой степенью сердечного заболевания. Детализированный анализ этой группы позволяет сделать следующие выводы. Диапазон возрастов (age) от 35 до 77 лет с медианным значением 60 лет. Большинство пациентов относится к возрастной категории старше 50 лет, что согласуется с увеличением риска сердечных заболеваний с возрастом. Группа включает значительно больше мужчин (99) по сравнению с женщинами (8), что может указывать на более высокую распространенность тяжелых сердечных заболеваний среди мужчин в этой выборке. Большинство пациентов (83 из 107) описывают боли (cp) как бессимптомные, что может свидетельствовать о высокой степени ишемии и серьезности сердечных проблем. Колебания уровня холестерина (chol) от 0 до 369 мг/дл с медианным значением 214 мг/дл. Значения максимального пульса (thalch) от 63 до 173 уд/мин, с медианой 122 уд/мин, могут указывать на ограниченную способность сердца к нагрузке. Наличие стенокардии при нагрузке (exang) у большинства пациентов (64 из 107) подтверждает серьезность сердечных проблем в этой группе. Значения депрессии ST-сегмента от 0 до 6.2, с медианой 1.0, подчеркивают значительные сердечные изменения при нагрузках. Электрокардиограмма (restecg) и другие показатели также подчеркивают различные состояния сердечной нагрузки и функции сердечной мышцы среди пациентов этой группы.

Группа с num = 3 включает пациентов с тяжелой степенью сердечного заболевания, характеризующейся серьезными симптомами, значительными изменениями в клинических параметрах и высоким уровнем стенокардии при нагрузке. Эти данные подчеркивают необходимость тщательного мониторинга и лечения для улучшения состояния и качества жизни пациентов с серьезными сердечными заболеваниями.

Группа num = 4 охватывает пациентов с наиболее тяжелой формой сердечного заболевания. Анализируя данные, можно сделать следующие выводы. Возраст пациентов варьируется от 38 до 77 лет, с медианой и средним возрастом около 59 лет, что подчеркивает преобладание сердечных заболеваний в старшей возрастной группе. Преобладают мужчины (26) по сравнению с женщинами (2), что указывает на высокую распространенность сердечных заболеваний среди мужской части населения. Большинство пациентов (23 из 28) испытывают бессимптомные боли, что может свидетельствовать о тяжелой степени ишемической болезни сердца. Значения артериального давления от 104 до 190 мм рт.ст., с медианой в 132 мм рт. ст., характерны для гипертонических состояний. Уровень холестерина (chol) варьируется от 0 до 407 мг/дл, с медианой 224 мг/дл. Максимальный пульс (thalch) от 84 до 182 уд/мин, с медианой 126.5 уд/мин, что может указывать на ограниченную способность к физической нагрузке и низкую эффективность сердечно-сосудистой системы. У большинства пациентов (15 из 28) наблюдается стенокардия при нагрузке (exang), что является ярким показателем наличия сердечных проблем. Значения депресии ST-сегмента от 0 до 4.4, с медианой 2.45, подчеркивают значительные изменения под нагрузкой, что коррелирует с тяжелой степенью ишемии. Наличие гипертрофии левого желудочка у 13 пациентов и другие изменения на ЭКГ подтверждают серьезные сердечные нарушения.

Группа с num = 4 представляет пациентов с наиболее тяжелым сердечным заболеванием. Характерные симптомы, высокие значения давления, изменения ЭКГ и депрессия ST-сегмента указывают на необходимость комплексного подхода к лечению и возможно, рассмотрение оперативных вмешательств для улучшения качества жизни и профилактики осложнений.

Обобщая данные по пяти группам пациентов, можно сделать следующие выводы:

1. Группа с num = 0 характеризуется отсутствием клинически значимых сердечных заболеваний, что подчеркивает важность профилактических мер и мониторинга здоровья даже у лиц без явных признаков заболеваний. Наличие повышенных значений холестерина и артериального давления в этой группе требует внимания в рамках ранней диагностики и профилактики сердечно-сосудистых заболеваний.

2. Группа с num = 1 демонстрирует начальные стадии сердечных заболеваний с характерными изменениями ЭКГ и признаками стенокардии. Результаты из этой группы могут служить основой для разработки ранних вмешательств и улучшения терапевтических стратегий.

3. Группа с num = 2 представляет пациентов с умеренной степенью сердечных нарушений. Наблюдаемые симптомы и клинические данные указывают на необходимость дополнительных исследований для уточнения диагностики и подбора лечения.

4. Группа с num = 3 включает пациентов с тяжелыми сердечными заболеваниями. Эта группа требует особого внимания, так как выраженные клинические проявления и серьезные изменения в показателях здоровья требуют интенсивного лечения и возможно, хирургического вмешательства.

5. Группа с num = 4 содержит пациентов с критическими формами сердечных заболеваний. Анализ показывает наличие симптомов, требующих неотложной медицинской помощи и, возможно, разработки новых подходов к лечению и реабилитации.

Построение модели линейной регрессии

Перед построением модели линейной регрессии к данным было применено логарифмирование для более эффективного статистического анализа и обеспечения более надежных и интерпретируемых результатов.

Результаты диагностики модели представлены следующими диаграммами (рис. 13).

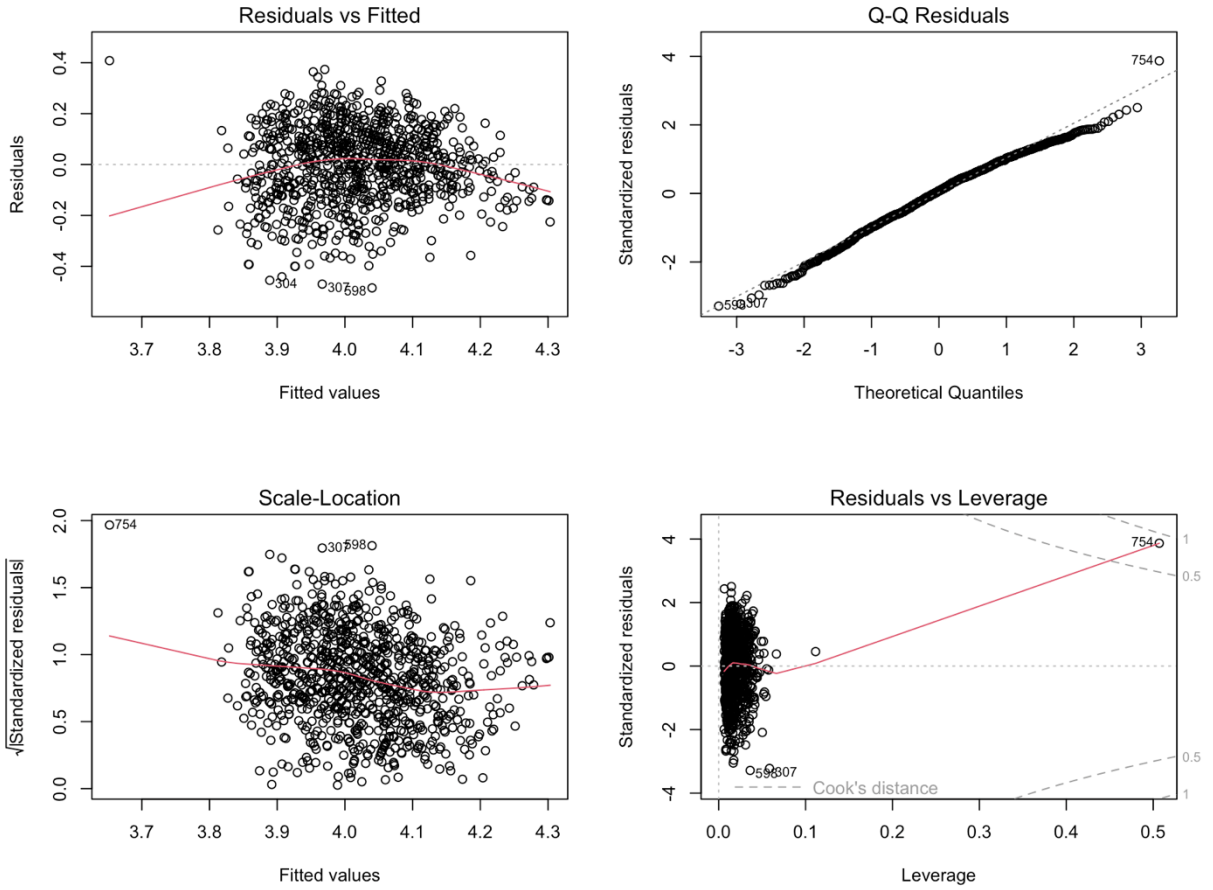

Рисунок 13. Диаграммы диагностики результатов линейной регрессии

На предоставленных графиках можно заметить несколько положительных аспектов:

1. Диаграмма “Residuals vs Fitted”. Хотя и виден определённый паттерн, большинство остатков сгруппированы около горизонтальной линии, что указывает на то, что для многих наблюдений модель предоставляет адекватные прогнозы.

2. Диаграмма “Q-Q Plot of Residuals”. Несмотря на некоторые отклонения на концах, большая часть точек следует прямой линии, что говорит о том, что остатки приближаются к нормальному распределению.

3. Диаграмма “Scale-Location”. Несмотря на некоторый тренд, который виден на графике, остатки не показывают экстремально большой дисперсии, что может указывать на то, что гетероскедастичность не слишком выражена.

4. Диаграмма “Residuals vs Leverage”. Большинство точек имеют низкое влияние (leverage), что говорит о том, что модель не доминируется небольшим количеством точек данных. Это положительный аспект, так как он указывает на стабильность оценок модели.

Данная модель линейной регрессии построена на логарифмированных данных, что было сделано для стабилизации дисперсии и уменьшения влияния выбросов. Судя по результатам, модель включает как значимые, так и незначимые предикторы (рис. 14):

1. (Intercept), или свободный член, значим и указывает на базовое значение зависимой переменной при нулевых значениях всех независимых переменных.

2. sexMale имеет отрицательный коэффициент, что предполагает, что значения целевой переменной для мужчин в среднем ниже, чем для женщин, однако это различие не является статистически значимым ($p > 0.05$).

3. Тип боли в груди, cp, имеет разные коэффициенты для разных категорий, но ни одна из них не показывает статистической значимости.

4. Артериальное давление в покое, trestbps, имеет положительный и значимый коэффициент, что указывает на важность этой переменной в модели.

5. Наличие высокого уровня сахара в крови натощак, fbsTRUE, также значима и имеет положительный коэффициент.

6. Нормальный результат электрокардиограммы в покое, restecgnormal, имеет отрицательный и статистически значимый коэффициент, что может указывать на негативное влияние нормальных результатов ЭКГ на целевую переменную.

7. Максимальная ЧСС, thalch, имеет отрицательный и значимый коэффициент, что говорит о том, что с увеличением максимального пульса целевая переменная уменьшается.

8. Депрессия ST-сегмента, oldpeak, также значима и имеет положительный коэффициент, указывающий на её влияние на целевую переменную.

Показатель R-squared говорит о том, что около 27.69% вариабельности данных объясняется моделью. F-статистика и соответствующее ей p-значение указывают на то, что модель в целом значима.

В то же время присутствуют переменные, которые не показывают статистической значимости и могут не вносить значимого вклада в модель.

```
> summary(model)

Call:
lm(data = logged_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48473 -0.09759  0.01123  0.10564  0.40813

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.830490   0.215145  22.452  < 2e-16 ***
sexMale                   -0.023812   0.012844  -1.854 0.064071 .
cpatypical angina         -0.014803   0.015568  -0.951 0.341954
cpnon-anginal              0.021807   0.013635   1.599 0.110094
cptypical angina           0.024870   0.024137   1.030 0.303103
trestbps                   0.100109   0.028310   3.536 0.000427 ***
chol                      -0.004250   0.003283  -1.295 0.195745
fbsTRUE                    0.077281   0.014436   5.353 1.10e-07 ***
restecgnormal             -0.059439   0.013486  -4.407 1.17e-05 ***
restecgst-t abnormality   -0.015927   0.016827  -0.947 0.344135
thalch                    -0.275639   0.030956  -8.904  < 2e-16 ***
exangTRUE                 -0.014004   0.012547  -1.116 0.264686
oldpeak                    0.081103   0.022644   3.582 0.000360 ***
slopeflat                 -0.034231   0.020921  -1.636 0.102145
slopeupsloping             0.009397   0.023203   0.405 0.685574
thalnormal                 0.016890   0.023303   0.725 0.468759
thalreversable defect      0.029054   0.024940   1.165 0.244348
num                        0.022918   0.005654   4.053 5.49e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1503 on 902 degrees of freedom
Multiple R-squared:  0.2769,    Adjusted R-squared:  0.2632
F-statistic: 20.32 on 17 and 902 DF,  p-value: < 2.2e-16
```

Рисунок 14. Результаты модели линейной регрессии

Кластерный анализ

Для проведения кластерного анализа набор данных, содержащий категориальные переменные, заранее преобразованные в числовые, был масштабирован с целью улучшения сравнимости переменных и повышения эффективности анализа.

Для разделения данных на 5 кластеров по значениям целевой переменной num (0, 1, 2, 3, 4) использована функция kmeans(). Результаты кластеризации представлены на рис. 15.

```
> print(kmeans_result)
K-means clustering with 5 clusters of sizes 199, 263, 137, 162, 159

Cluster means:
         age   trestbps       chol     thalch    oldpeak        num sex_numeric cp_numeric fbs_numeric
1  0.4619202  0.3945917  0.5050543 -0.5547243  0.5751608  0.4259759   0.3812177 -0.5214490  0.52253392
2 -0.5536624 -0.1462810  0.3212272  0.5898662 -0.5473438 -0.6716721   0.5166500  0.5770026 -0.15378743
3  0.3152235  0.1620346  0.2111663 -0.1459717  0.8120117  0.8725425   0.2305078 -0.5424917 -0.03166794
4  0.3204345 -0.2075185 -1.7375072 -0.6292254 -0.1150981  0.5710158   0.3956577 -0.3845445 -0.17796351
5 -0.2604083 -0.1800792  0.4248920  0.4854590 -0.3968887 -0.7557379  -1.9334426  0.5574481 -0.19100352
  restecg_numeric exang_numeric  slope_num    thal_num
1      0.38926693    0.72037987 -0.4604611 -0.53609502
2     -0.09283408   -0.66522753  0.2651118 -0.21026456
3     -0.81626007    0.70889025 -0.1110129  1.48340685
4      0.50394532    0.07246091 -0.1274289  0.02925349
5     -0.14377526   -0.48589550  0.3632680 -0.28920324
```

Рисунок 15. Результаты кластеризации

Подробный анализ кластеров, полученных с помощью k-средних, позволяет глубже понять различия в данных:

1. Кластер 1. Этот кластер характеризуется относительно высокими значениями возраста, кровяного давления, уровня холестерина и степени депрессии ST-сегмента, но низкими значениями максимального сердцебиения. Пациенты в этой группе имеют повышенный средний уровень риска сердечных заболеваний, что подчеркивается значением целевой переменной num.
2. Кластер 2. Этот кластер имеет отрицательные значения по возрасту и кровяному давлению, но положительные значения по уровню холестерина и максимального сердцебиения. Степень депрессии ST-сегмента низкая. Пациенты в этом кластере имеют наименьший

средний риск сердечных заболеваний, что отражается в значениях целевой переменной num.

3. Кластер 3. Эта группа имеет умеренные средние значения по возрасту, кровяному давлению и уровню холестерина, но значительно выше среднего значения степени депрессии ST-сегмента и целевой переменной num, что указывает на более высокий риск сердечных заболеваний.
4. Кластер 4. Пациенты здесь имеют средние значения возраста и негативные значения уровня холестерина, что может свидетельствовать о неполных или некорректных данных, а также самые низкие значения максимального сердцебиения. Средний риск сердечных заболеваний num в этой группе выше среднего.
5. Кластер 5. Этот кластер отмечен отрицательными значениями по возрасту и кровяному давлению, положительными по уровню холестерина и максимального сердцебиения, но низкой степенью депрессии ST-сегмента. Пациенты здесь имеют самый низкий риск сердечных заболеваний num, таким образом, это наиболее здоровая группа среди всех.

Заключение

Исследование на основе анализа данных “Heart Disease (UCI)” подтверждает предположение о том, что определённые медицинские показатели могут быть связаны с риском развития сердечных заболеваний. Результаты распределения ключевых переменных показали значительное варьирование в значениях артериального давления, уровня холестерина и сердечного ритма среди пациентов, что отражено в гистограммах и боксплотах. Эти визуализации подчеркнули различия между группами пациентов и обозначили специфические риски для отдельных категорий.

Анализ корреляций выявил значимые статистические связи между такими параметрами, как возраст, уровень холестерина, масимальная частота сердечных сокращений, депрессия ST-сегмента, пол, тип болей в груди, тип болей при физической нагрузке и наличием сердечных заболеваний, подтверждая первоначальную гипотезу исследования. Кластерный анализ разграничил пациентов на группы с различными уровнями риска и клиническими профилями, что может быть использовано для более целенаправленного подхода в профилактике и лечении.

Результаты линейной регрессии позволили количественно оценить влияние каждого фактора на вероятность развития сердечного заболевания. В частности, показано, что такие факторы, как возраст, пол, уровень холестерина и артериальное давление, имеют значимое влияние на сердечное здоровье.

В целом, данные подтверждают необходимость комплексного подхода к анализу риска сердечных заболеваний, включая использование статистического анализа для определения важнейших факторов риска. Представленные результаты могут быть полезны для разработки персонализированных медицинских рекомендаций и улучшения стратегий лечения и профилактики сердечных заболеваний.

Список использованных источников:

1. Janosi, Andras, Steinbrunn, William, Pfisterer, Matthias, and Detrano, Robert. (1988). Heart Disease. UCI Machine Learning Repository [Электронный ресурс]. URL: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data
2. Программный код [Электронный ресурс]. URL: https://github.com/Vejber/Data_science/blob/main/final/Heart.R.