Федеральное государственное автономное образовательное учреждение высшего образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет экономических наук Образовательная программа: Экономика

Название дисциплины: Эконометрика 1

Преподаватель семинаров: Бывальцева-Станкевич Анастасия Александровна; (аспирант, преподаватель: Факультет экономических наук / Департамент прикладной экономики, менеджер: Факультет экономических наук / Отдел сопровождения проектной работы)

Отчет о выполнении модульного домашнего задания №2

«Эконометрическое исследование по влиянию физико-химического состава вина на его качество»

Выполнили студенты:

Монтлевич Дмитрий Константинович,

Князева Мария Александровна,

Кадымова Зарина Суфияновна

Москва, 2024г.

Оглавление

Введение

Цель исследовательской работы: определить и описать влияние физико-химических факторов на органолептические свойства португальского вина "Vinho Verde", представленные через показатель качества (оценка от 0 до 10).

Задачи исследования:

1. Разработать экономическую модель:
    - Сформировать список объясняющих переменных и предположить как они могут влиять на таргет..
    - Сформулировать содержательные гипотезы для проверки в рамках исследования.
2. Провести анализ данных:
    - Проверить наличие пропущенных данных, выбросов и других аномалий
    - Изучить распределение физико-химических характеристик вина и их связь с показателем качества.
3. Моделирование взаимосвязей:
    - Построить эконометрическую модель (множественную линейную регрессию) для описания влияния факторов на качество вина.
    - Содержательно интерпретировать результаты оценивания моделей, включая экономический смысл коэффициентов.
4. Сделать выводы:
    - Сделать выводы относительно сформулированных гипотез
    - Сделать общие выводы относительно влияния физико-химического состава вина на его качество

Актуальность:

Качество вина — это важный показатель, который влияет на его потребительскую привлекательность и рыночную стоимость. Изучение влияния физических и химических параметров на качество вина:

- Позволяет понять, какие свойства наиболее значимы для достижения высокого качества.
- Может помочь производителям оптимизировать технологический процесс, улучшить органолептические характеристики и повысить ценность продукции.
- Является примером применения методов анализа данных в реальных задачах агроиндустрии и пищевой промышленности.

Данное исследование также актуально в контексте использования современных эконометрических и аналитических инструментов для решения задач в области анализа больших данных.

Винодельческая отрасль занимает особое место в мировой экономике, объединяя в себе традиции и инновации. Совершенствование качества вина требует глубокого изучения его характеристик и разработки более точных методов сертификации. Это включает объективные физико-химические измерения и субъективные сенсорные тесты, которые помогают оценить восприятие вкуса, аромата и текстуры напитка. Такой подход позволяет учитывать широкий спектр факторов, влияющих на качество продукции.

Процедуры сертификации качества вина базируются на сочетании данных лабораторных исследований и мнений экспертов-дегустаторов. Ключевые физико-химические показатели, такие как кислотность, содержание алкоголя, сахар и плотность, позволяют определить базовые свойства вина. Сенсорные тесты, напротив, оценивают более тонкие аспекты, связанные с вкусовыми и ароматическими ощущениями. Совокупность этих методов формирует комплексное понимание качества напитка.

Современные исследования всё больше акцентируют внимание на изучении химических соединений, отвечающих за уникальные органолептические свойства вина. Большой интерес представляют летучие компоненты, связанные с сортовыми особенностями винограда. При этом важно учитывать, что разные соединения имеют различный сенсорный порог. Например, вещества с низкой концентрацией могут оказывать сильное влияние на вкус, тогда как высококонцентрированные соединения могут иметь минимальное воздействие. Это создаёт дополнительные сложности в анализе и требует разработки гибких аналитических методов.

Таким образом, изучение взаимосвязей между химическим составом и вкусовыми характеристиками вина играет ключевую роль в повышении объективности оценки качества. Это направление позволяет не только совершенствовать методы сертификации, но и раскрывать потенциал каждого сорта винограда, подчеркивая его уникальность.

Описание данных:

Наши данные были взяты с платформы Kaggle, где содержатся множество датасетов, подходящих для нашего исследования. В частности был выбрал датасет **Wine Quality,** описывающий различные физико-химические характеристики португальского красного и белого вина **Vinho Verde** из одноименного региона. Данные были собраны с мая 2004 по февраль 2007. Датасет предоставлен официальной организацией сертификации и контроля качества CVRVV, которая занимается поддержанием стандартов производства и маркетинга данного вина.

При сборе данных была использована автоматизированная система контроля iLab. Эта система отвечает за координацию тестирования винных образцов, начиная от подачи запросов производителей и заканчивая лабораторными и дегустационными процедурами. Каждый образец вина подвергался как физико-химическому анализу, так и экспертной сенсорной оценке. [1]

Датасет включает 1 599 записей красного вина и 4 898 записей белого вина (до предобработки). Сенсорная оценка проводилась в формате слепой дегустации, где как минимум три профессиональных эксперта выставляли свои баллы по шкале от 0 до 10. Итоговая оценка качества определялась как медианное значение среди выставленных оценок. Эти данные служат основой для анализа, направленного на изучение факторов, влияющих на восприятие качества винной продукции, и позволяют выявить взаимосвязи между физико-химическими характеристиками и экспертными оценками.

Экономическая модель

a. Описание переменных:

Независимые переменные (физико-химические характеристики):

1. **Тип вина**: Категориальная переменная, указывающая тип вина (1 — красное, 0 — белое).
2. **Фиксированная кислотность (г/дм^3)**: Сумма органических кислот, которые остаются в вине после завершения процесса ферментации (например, винная, яблочная кислоты). Влияет на вкус вина.
3. **Летучая кислотность (г/дм^3)**: Количество кислот, склонных к испарению, таких как уксусная кислота. Высокая летучая кислотность может негативно сказываться на вкусе, так как связана с дефектами производства.
4. **Лимонная кислота (г/дм^3)**: Природная кислота, добавляющая свежести вкусу вина. Обычно присутствует в небольших количествах и может улучшать общий баланс кислотности.
5. **Остаточный сахар (г/дм^3)**: Количество сахара, оставшегося в вине после окончания ферментации. Этот параметр особенно важен для сладких и полусладких вин, а также для сбалансирования кислотности.
6. **Хлориды (г/дм^3)**: Концентрация хлоридов, влияющих на соленость и общую текстуру вина. Высокий уровень может негативно сказываться на вкусе, делая его не гармоничным.
7. **Свободный диоксид серы (мг/дм^3)**: Концентрация свободного диоксида серы SO_2, добавляемого в вино для предотвращения окисления и сохранения свежести. Важный показатель для сохранения качества вина в процессе хранения.
8. **Общий диоксид серы (мг/дм^3)**: Суммарное количество диоксида серы SO_2, включая связанный и свободный SO_2. Показатель качества и устойчивости вина к окислению.
9. **Плотность (г/см^3)**: Зависит от содержания сахара и спирта в вине. Чем выше содержание сахара или алкоголя, тем больше плотность вина.
10. **pH (измеряется в в единицах pH)**: Уровень кислотности вина. Влияет на стабильность продукта, а также на восприятие вкуса. Более низкий pH делает вино более устойчивым к микробиологическим рискам.
11. **Сульфаты (г/дм^3)**: Вещество, способствующее улучшению антиоксидантных свойств и восприятия вкуса. Связывается с сохранением свежести и улучшением структуры вина.
12. **Спирт (доля алкоголя в %)**: Концентрация этилового спирта в вине, выраженная в процентах. Один из ключевых показателей, влияющих на текстуру, вкус и восприятие крепости вина.

Зависимая переменная:

13. **Качество**: Итоговая органолептическая оценка вина, выставляемая экспертами на основе слепой дегустации. Значения представлены в диапазоне от 0 (очень плохо) до 10 (превосходно). Медианное значение из оценок нескольких дегустаторов отражает восприятие качества продукта. Описание было составлено на основе статьи:MODELING THE PREFERENCE OF WINE QUALITY USING LOGISTIC REGRESSION TECHNIQUES BASED ON PHYSICOCHEMICAL PROPERTIES.[4]

b. Выбор переменных:

Данный набор объясняющих переменных был выбран из-за их прямой связи с физико-химическими характеристиками вина, которые существенно влияют на его вкусовые и ароматические свойства. Эти переменные, такие как фиксированная и летучая кислотность, уровень алкоголя, остаточный сахар, хлориды и pH, играют ключевую роль в формировании вкусового баланса и стабильности вина. Например, кислотность отвечает за свежесть вкуса, а уровень остаточного сахара влияет на восприятие сладости.

Дополнительно, показатели, такие как содержание диоксида серы (свободного и общего), важны для оценки сохранности и предотвращения порчи продукта, а плотность и сульфаты связаны с текстурой и антиоксидантными свойствами[2].Таким образом, выбранные показатели обеспечивают комплексный подход к анализу качества вина и позволяют эффективно моделировать экспертные оценки.

c. Оценка влияния независимых переменных:

С точки зрения линейных моделей коэффициенты переменных могут иметь такое влияние:

Фиксированная кислотность: Коэффициент **b_1** ожидается положительным, так как сбалансированная кислотность обеспечивает свежесть и гармоничность вкуса, что способствует улучшению качества.

Летучая кислотность: b_2 предположительно отрицательный, поскольку высокое значение летучей кислотности связано с дефектами вина.

Лимонная кислота: b_3 должен быть положительным, так как лимонная кислота добавляет свежести и улучшает баланс вкуса.

Остаточный сахар: b_4 может быть положительным для белых вин (так как сладость повышает их привлекательность), но менее значительным или отрицательным для сухих красных вин.

Хлориды: b_5 предположительно отрицательный, поскольку высокий уровень хлоридов приводит к нежелательной солености и ухудшению вкуса.

Свободный и общий диоксид серы: b_6 и **b_7** ожидаются положительными при оптимальных уровнях, так как SO_2 предотвращает окисление и улучшает стабильность вкуса.

Плотность: b_8 может быть нейтральным или слабо положительным, так как плотность напрямую зависит от содержания сахара и алкоголя.

pH: b_9 предположительно отрицательный, поскольку низкий pH способствует улучшению стабильности и вкуса вина.

Сульфаты: b_10 ожидается положительным, так как сульфаты усиливают текстуру и антиоксидантные свойства вина.

Содержание алкоголя: b_11 должен быть положительным, поскольку высокий уровень алкоголя ассоциируется с насыщенностью вкуса, но чрезмерные значения могут снизить общую оценку.

d. Гипотезы:

Гипотеза 1:
Увеличение содержания алкоголя в вине положительно влияет на оценку качества.
Обоснование: Алкоголь является одним из ключевых факторов, определяющих насыщенность вкуса и ароматическую структуру вина. Более высокое содержание алкоголя часто ассоциируется с улучшением вкусового профиля, особенно для красных вин.

Гипотеза 2:
Высокая летучая кислотность отрицательно влияет на оценку качества вина.
Обоснование: В статье Correlating Wine Quality Indicators to Chemical and Sensory Measurements отмечается, что летучая кислотность указывает на наличие уксусной кислоты и других кислот, которые могут сигнализировать о дефектах производства или порче вина. Это негативно отражается на вкусовых свойствах и снижает общую оценку. [2]

Гипотеза 3:
Остаточный сахар положительно влияет на качество белых вин, но имеет слабую или отрицательную связь с качеством красных вин.
Обоснование: Исследователи в статье Measuring Wine Quality and Typicity обсуждают сахар, фактор влияющий на качество вина. На основании этого мы хотим проверить эту гипотезу. Для белых вин небольшой уровень остаточного сахара может улучшить восприятие вкуса, добавляя сбалансированную сладость. Однако для сухих красных вин сладость может восприниматься как недостаток и снижать оценку.[3]

Предварительный анализ данных

Предобработка данных:

Для начала мы провели анализ на наличие пропусков в данных. Пропущенных значений оказалось достаточно мало (38) относительно общего объёма наблюдений, так что мы решили их удалить.
В данных обнаружилось достаточно большое количество дубликатов (1168). Было принято решение их тоже удалить, так как они не несут никакой новой информации и соответственно не помогут нам в исследовании зависимостей между признаками.
Также мы закодировали категориальную переменную в дамми-переменную, приняв значение 'red' = 1, а значение 'white' = 0.

Распределение типов вин по оценкам экспертов:

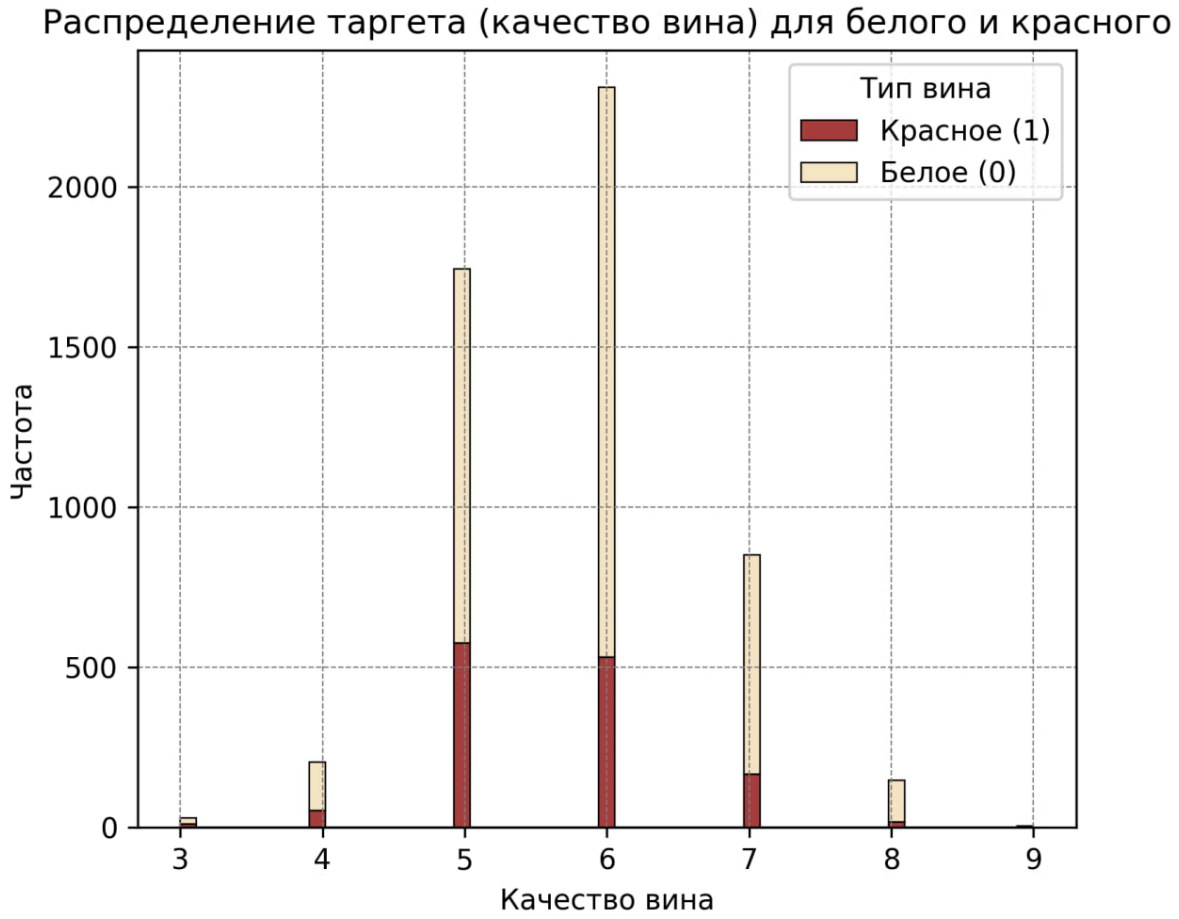


Рис.1 Распределение таргета (качество) для белого и красного вина

quality	Белое вино	Красное вино
3	0.005074	0.007391

4	0.038559	0.038433
5	0.296550	0.425721
6	0.451547	0.392461
7	0.173770	0.123429
8	0.033232	0.012565
9	0.001268	0.000000

Таблица 1. Относительные частоты белого и красного вина

По Рис. 1 и Таблице 1 заметно, что качество большинства вин (как белого, так и красного) сосредоточено в диапазоне 5–7. Однако белое вино имеет большую долю высоких оценок (7 и выше) относительно своего количества, что может указывать на более высокое качество белого вина в выборке. Красное вино распределено более равномерно по качеству, но с преобладанием среднего уровня (5–6).

Анализ переменных на выбросы:

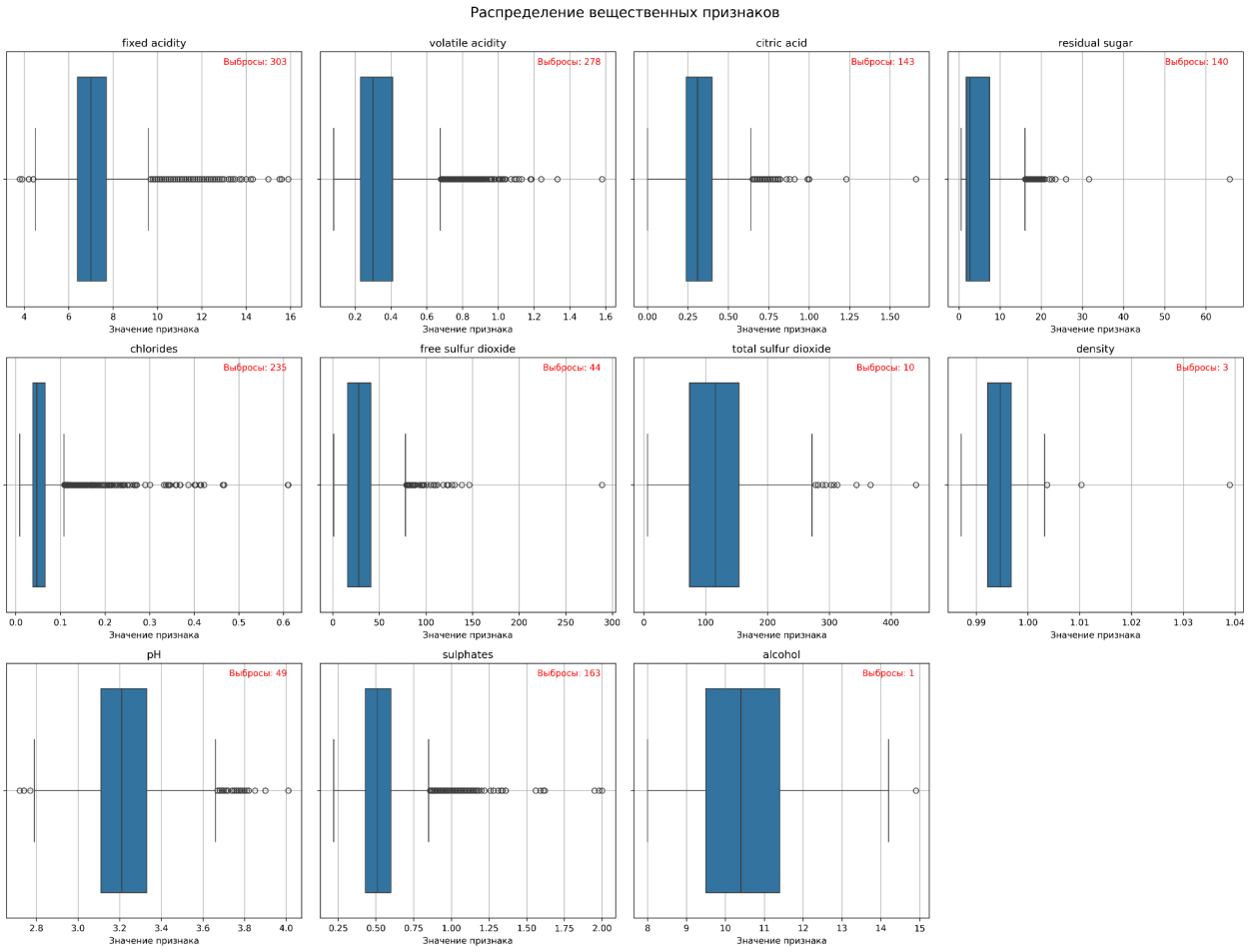


Рис.2 Распределение вещественных признаков

На представленных графиках (Рис. 2) видны box-plot наших переменных.Выбросы есть, но их не очень много относительно общего количества данных. Больше всего выбросов в признаках: fixed acidity (303), volatile acidity (278) и chlorides (235). Они могут быть связаны с особенностями сырья, технологиями производства (например, обработка SO_2, ферментация), различиями между типами вин (сухие, сладкие, красные, белые), а также возможными ошибками в данных или измерениях.

Анализ распределения переменных:

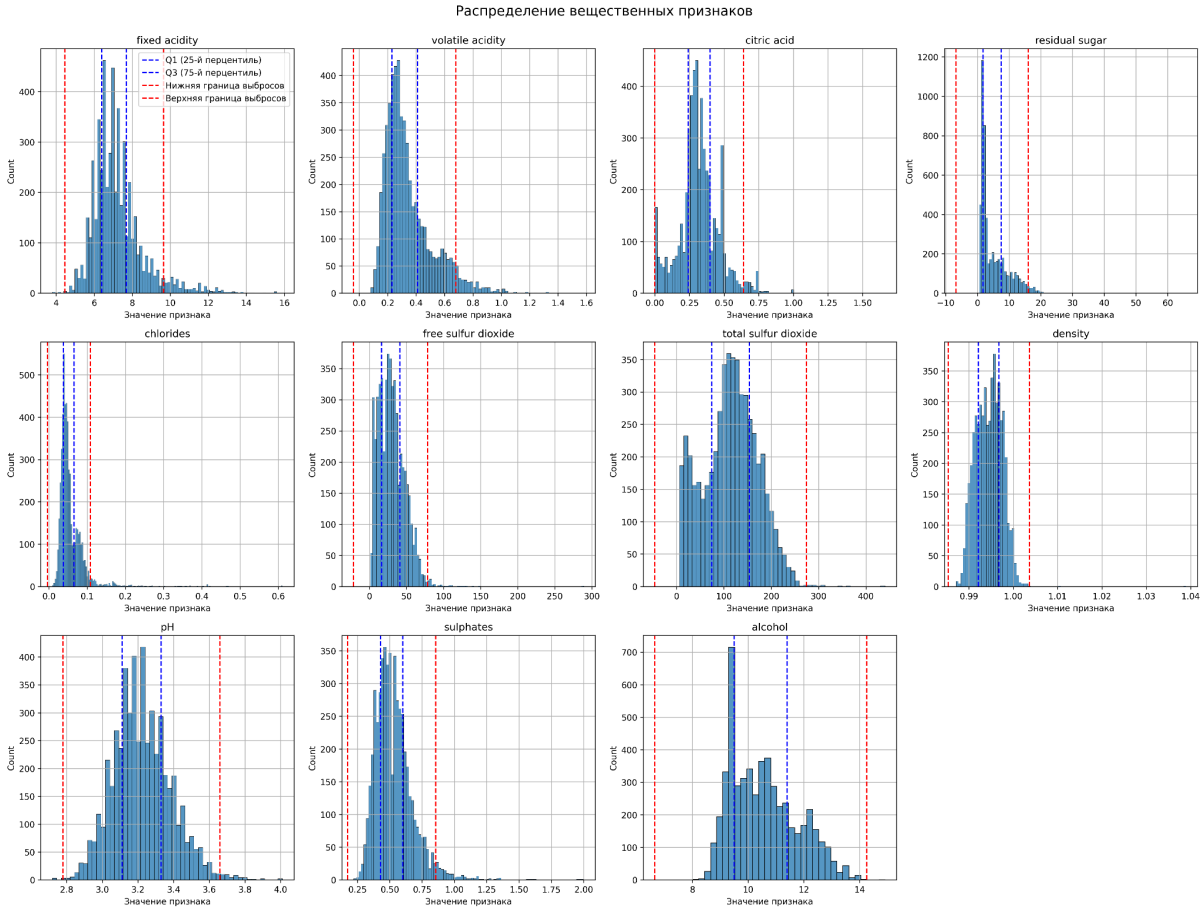


Рис.3 Гистограммы распределения переменных.

По Рис.3 заметно, что распределение признака pH больше всего похоже на нормальное. Все остальные графики не сильно схожи с гистограммами нормального распределения. Почти у всех признаков можно заметить правостороннюю асимметрию, именно поэтому выбросов больше со стороны возрастания признака.

Дополнительно был проведен тест Смирнова-Колмогорова на соответствие распределения нормальному на уровне значимости 5%. По результатам теста гипотеза

о принадлежности признаков нормальному распределению отвергается на любом разумном уровне значимости (в том числе и на уровне 5%).

Корреляционный анализ:

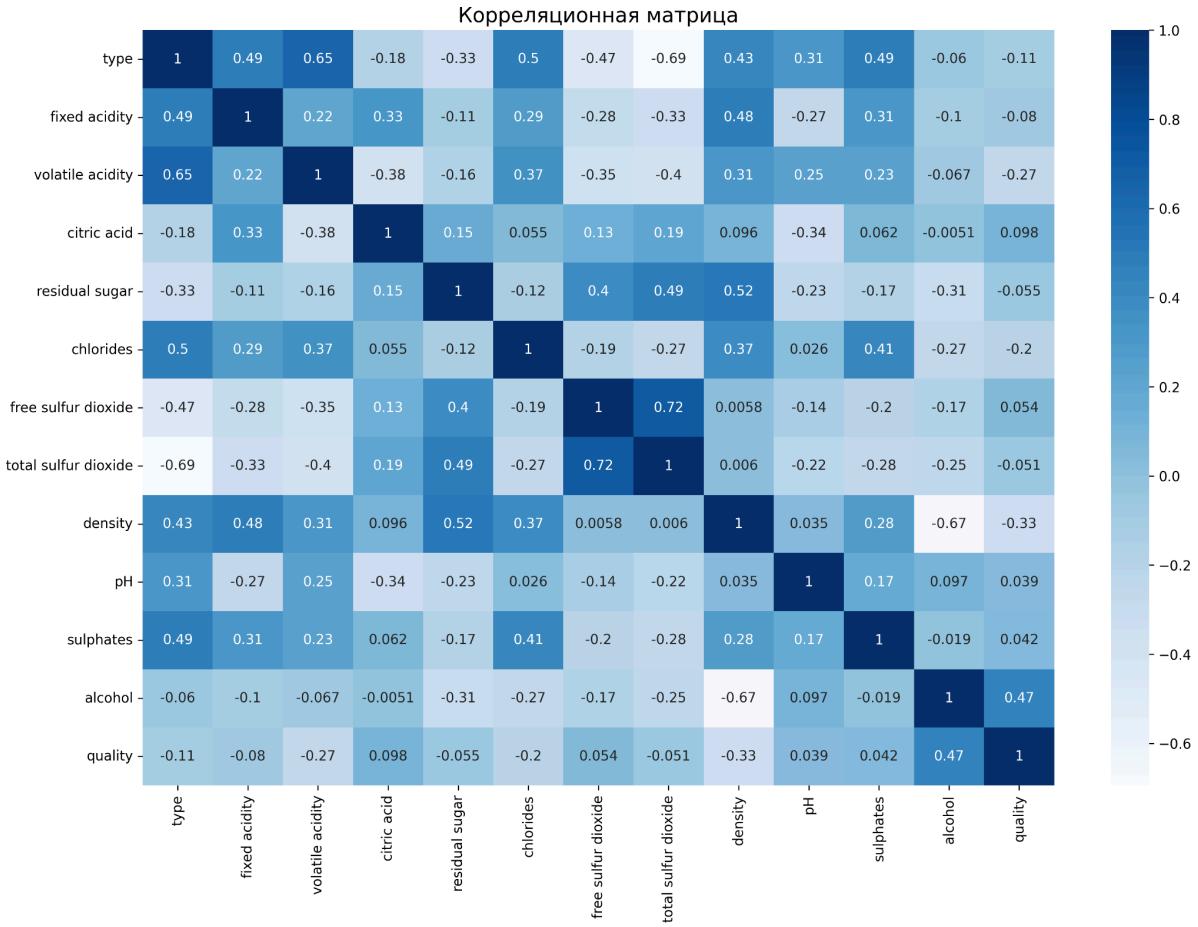


Рис.4 Корреляционная матрица признаков

На основе корреляционной матрицы (Рис. 4) можно сделать выводы о взаимосвязи признаков:

- total sulfur dioxide и free sulfur dioxide: r = 0.72 — сильная корреляция. Это ожидаемо, так как эти признаки связаны химически (общий диоксид серы включает свободный диоксид серы).
- density и residual sugar: r = 0.52 — умеренно высокая корреляция. В целом, в описании переменных мы предполагали, то будет взаимосвязь данных признаков.
- type и volatile acidity: r = 0.65 — высокая связь.

В целом, корреляции между большинством признаков и "quality" слабые, что может означать сложный многомерный характер влияния признаков на качество.

Анализ однородности выборок (и немного корреляции):

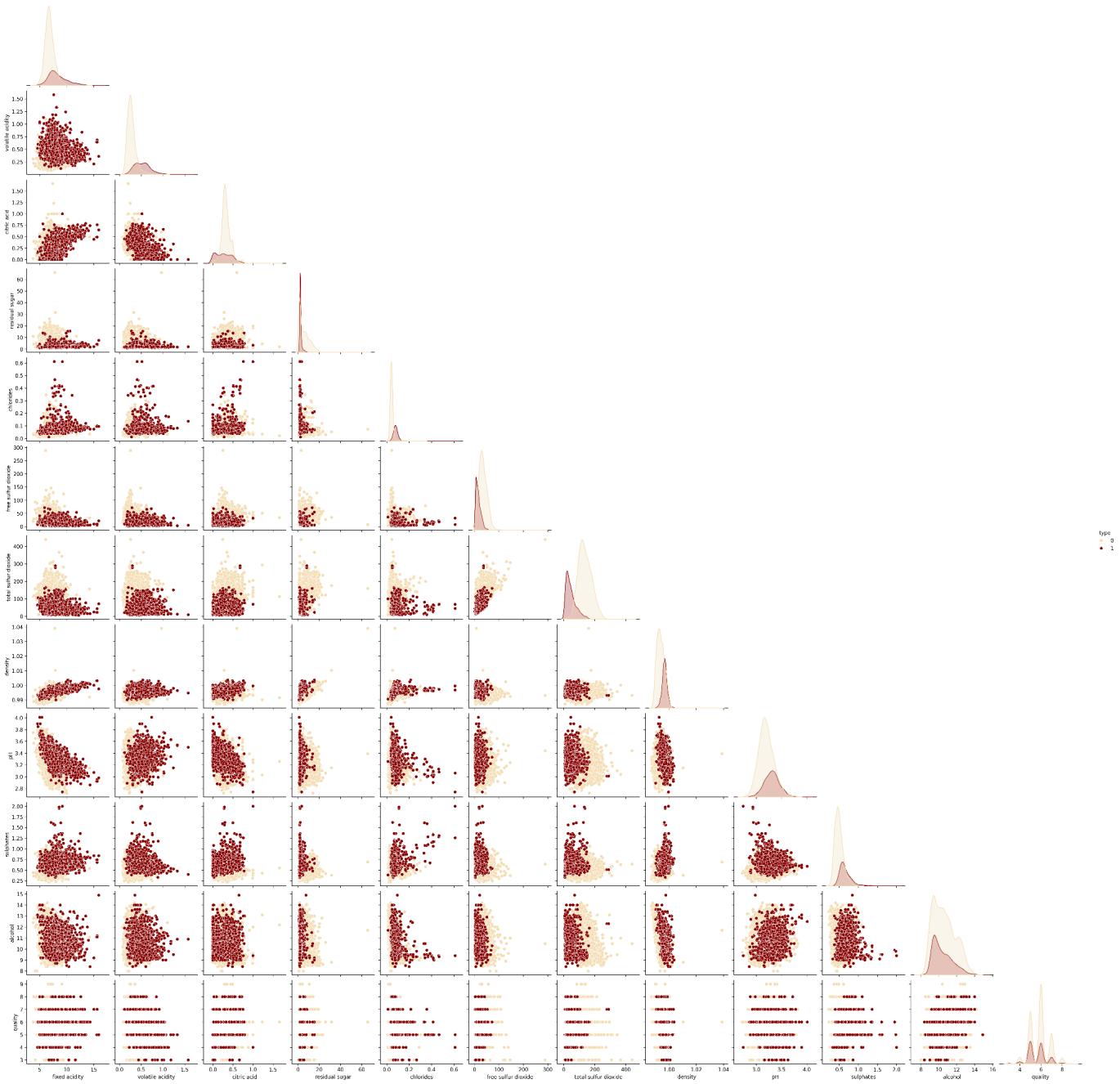


Рис.5 Парные диаграммы рассеяния для объясняющих переменных.

- Действительно заметна корреляция между признаками: density и residual sugar - вина с более высоким содержанием остаточного сахара имеют большую плотность. free sulfur dioxide и total sulfur dioxide - с ростом свободного диоксида серы увеличивается общее содержание серы (что логично).
- Белое и красное вино разделяются по признакам: volatile acidity (летучая кислотность): красное вино имеет более высокие значения по сравнению с

белым. citric acid (лимонная кислота): белое вино характеризуется более высокими значениями. residual sugar (остаточный сахар): у белого вина остаточный сахар чаще выше, чем у красного. density (плотность): белое вино чаще имеет большую плотность, что связано с остаточным сахаром.

Тест Чоу:

Проверим с помощью теста Чоу, действительно ли надо разделять красное и белое вина при построении модели или их общая модель хорошо описывает обе группы:

H_0 (нулевая гипотеза): Модель одинаково хорошо описывает обе группы.

H_1 (альтернативная гипотеза): Каждая группа лучше описывается своей собственной моделью.

Результаты теста:

статистика = 7.7576: Это значение говорит о том, что модели для двух подгрупп (красное и белое вино) значительно улучшают объяснение данных по сравнению с объединённой моделью.

p-value = 0.0000 (очень малое значение): Это значение указывает на статистически значимые различия между моделями для красного и белого вина. Нулевая гипотеза (о том, что объединенная модель так же хороша, как отдельные - отвергается.

Оценка моделей

Тест Чоу показал, что раздельные модели для красного и белого вина будут лучше объяснять данные, чем их совместная модель. Поэтому мы разбили данные относительно признака type и построили отдельные модели для красного и белого вина.
Перед обучением моделей линейной регрессии все вещественные признаки были предварительно отмасштабировать через функцию *StandartScaler.* Через МНК были выведены оценки получившихся моделей:

Модель для красного вина:

Dep. Variable:	*quality*	*R-squared:*	*0.365*
Model:	*OLS*	*Adj. R-squared:*	*0.359*
Method:	*Least Squares*	*F-statistic:*	*69.97*
Date:	*Fri, 13 Dec 2024*	*Prob (F-statistic):*	*9.2E-124*
Time:	*13:25:16*	*Log-Likelihood:*	*-1350.2*
No. Observations:	*1353*	*AIC:*	*2724*
Df Residuals:	*1341*	*BIC:*	*2787*
Df Model:	*11*		
Covariance Type:	*nonrobust*		

Таблица 2. Результаты оценки регрессии для красного вина

	coef	*std err*	*t*	*P>|t|*	*[0.025*	*0.975]*
const	5,7313	0,069	82,69	0	5,595	5,867
fixed acidity	0,0136	0,038	0,356	0,722	-0,061	0,089
volatile acidity	-0,1895	0,022	-8,625	0	-0,233	-0,146
citric acid	-0,0254	0,024	-1,066	0,287	-0,072	0,021
residual sugar	0,0336	0,076	0,44	0,66	-0,116	0,183
chlorides	-0,0702	0,017	-4,238	0	-0,103	-0,038
free sulfur dioxide	0,0599	0,043	1,396	0,163	-0,024	0,144
total sulfur dioxide	-0,1548	0,045	-3,413	0,001	-0,244	-0,066
density	-0,0203	0,071	-0,285	0,776	-0,16	0,12
pH	-0,0737	0,034	-2,16	0,031	-0,141	-0,007
sulphates	0,1369	0,019	7,192	0	0,1	0,174
alcohol	0,3469	0,035	9,942	0	0,278	0,415

Таблица 3. Результаты оценки коэффициентов регрессии для красного вина

цветом выделены не значимые признаки на уровне 5%

Omnibus:	25,511	Durbin-Watson:	1,782
Prob(Omnibus):	0	Jarque-Bera (JB):	36,328
Skew:	-0,2	Prob(JB):	1,29E-08
Kurtosis:	3,696	Cond. No.	18,2

Таблица 4. Результаты оценки регрессии для красного вина

Выводы:

- Коэффициент детерминации (R^2) = 0.365, что указывает на то, что модель объясняет около 36.5% вариации в качестве вина. Это умеренный уровень объяснения, предполагающий, что есть другие факторы, не учтенные в модели.
- Значение F-статистики (69.97) и очень низкое p-value (<0.0001) указывают на то, что модель статистически значима в целом.
- Переменные fixed acidity, citric acid, residual sugar, free sulfur dioxide и density не оказали статистически значимого влияния на качество красного вина (p-value > 0.05). Все остальные переменные - статистически значимы для этой модели.

Модель для белого вина:

Dep. Variable:	***quality***	***R-squared:***	***0,301***
Model:	***OLS***	***Adj. R-squared:***	***0,299***
Method:	***Least Squares***	***F-statistic:***	***153,5***
Date:	***Fri, 13 Dec 2024***	***Prob (F-statistic):***	***5,2E-295***
Time:	***13:25:16***	***Log-Likelihood:***	***-4433,2***
No. Observations:	***3942***	***AIC:***	***8890***
Df Residuals:	***3930***	***BIC:***	***8966***
Df Model:	***11***		
Covariance Type:	***nonrobust***		

Таблица 5. Результаты оценки регрессии для белого вина

	coef	***std err***	***t***	***P>|t|***	***[0.025***	***0.975]***
const	5,6296	0,023	242,738	0	5,584	5,675
fixed acidity	0,0784	0,029	2,661	0,008	0,021	0,136
volatile acidity	-0,2698	0,021	-12,875	0	-0,311	-0,229
citric acid	0,0351	0,016	2,254	0,024	0,005	0,066
residual sugar	0,3292	0,036	9,03	0	0,258	0,401
chlorides	-0,0188	0,021	-0,89	0,373	-0,06	0,023

free sulfur dioxide	0,0883	0,017	5,307	0	0,056	0,121
total sulfur dioxide	-0,036	0,024	-1,523	0,128	-0,082	0,01
density	-0,4251	0,06	-7,089	0	-0,543	-0,308
pH	0,134	0,018	7,302	0	0,098	0,17
sulphates	0,096	0,017	5,77	0	0,063	0,129
alcohol	0,2425	0,031	7,887	0	0,182	0,303

Таблица 6. Результаты оценки коэффициентов регрессии для белого вина

цветом выделены не значимые признаки на уровне 5%

Omnibus:	115,8	Durbin-Watson:	1,771
Prob(Omnibus):	0	Jarque-Bera (JB):	287,16
Skew:	-0,076	Prob(JB):	4,41E-63
Kurtosis:	4,314	Cond. No.	11,5

Таблица 7. Результаты оценки регрессии для красного вина

Выводы:

- Коэффициент детерминации (R^2) = 0.301, что означает, что модель объясняет около 30.1% вариации качества белого вина. Качество этой модели хуже чем для красного вина, но в целом не сильно отличается.
- Значение F-статистики (153,5) и очень низкое p-value (<0.0001) указывают на то, что эта модель тоже статистически значима.
- В отличие от модели для красного вина, тут всего 2 статистически незначимых признака на уровне 5%: chlorides и total sulfur dioxide (p-value > 0.05). Все остальные признаки - значимы для этой модели.

Общая модель:

Dep. Variable:	***quality***	***R-squared:***	***0.312***
Model:	***OLS***	***Adj. R-squared:***	***0.311***
Method:	***Least Squares***	***F-statistic:***	***199.7***
Date:	***Fri, 13 Dec 2024***	***Prob (F-statistic):***	***0***
Time:	***13:25:16***	***Log-Likelihood:***	***-5845.3***
No. Observations:	***5295***	***AIC:***	***11720***
Df Residuals:	***5282***	***BIC:***	***11800***
Df Model:	***12***		
Covariance Type:	***nonrobust***		

Таблица 8. Результаты оценки регрессии для белого вина

	coef	***std err***	***t***	***P>|t|***	***[0.025***	***0.975]***

const	5,7131	0,018	308,928	0	5,677	5,749
type	0,3247	0,061	5,343	0	0,206	0,444
fixed acidity	0,1024	0,023	4,548	0	0,058	0,147
volatile acidity	-0,2264	0,015	-15,282	0	-0,255	-0,197
citric acid	0,0173	0,013	1,34	0,18	-0,008	0,043
residual sugar	0,2475	0,029	8,493	0	0,19	0,305
chlorides	-0,0336	0,013	-2,603	0,009	-0,059	-0,008
free sulfur dioxide	0,105	0,015	6,965	0	0,075	0,135
total sulfur dioxide	-0,0899	0,02	-4,459	0	-0,129	-0,05
density	-0,2957	0,046	-6,471	0	-0,385	-0,206
pH	0,0981	0,016	6,185	0	0,067	0,129
sulphates	0,1119	0,013	8,907	0	0,087	0,136
alcohol	0,2747	0,023	11,847	0	0,229	0,32

Таблица 9. Результаты оценки коэффициентов общей регрессии

цветом выделены не значимые признаки на уровне 5%

Omnibus:	148.05	Durbin-Watson:	1.775
Prob(Omnibus):	0	Jarque-Bera (JB):	339.038
Skew:	-0.12	Prob(JB):	2.39E-74
Kurtosis:	4.216	Cond. No.	13.3

Таблица 10. Результаты оценки регрессии общей модели

Выводы:

- Общая модель имеет $R^2 = 0.312$, что ниже, чем $R^2 = 0.365$ для модели красного вина, но чуть выше, чем $R^2 = 0.301$ для модели белого вина. Это говорит о том, что общая модель дает лишь усредненные оценки, которые действительно хуже описывают зависимости внутри каждой группы.
- Значение F-статистики (199.7) и очень низкое p-value (<0.0001) указывают на то, что эта модель тоже статистически значима.
- В отличие от раздельных моделей, тут всего 1 статистически незначимый признак на уровне 5%: citric acid (p-value > 0.05). Все остальные признаки - значимы для этой модели.

Заключение

Общие выводы

На основе теста Чоу и анализа оценок моделей стало ясно, что разделение данных на подгруппы (тип вина) помогает выявить различия между ними. Это позволяет учесть индивидуальные особенности каждого вида вина и получить более точное понимание влияния различных признаков на их качество.

Оценки качества всех трех моделей в целом схожи: они все статистически значимы и R^2 не сильно отличается. Однако по точности лучше всех оказалась модель для красного вина (R^2 = 0.365), значит наши физико-химические признаки лучше всего помогают объяснить качество именно красного вина.

Можно заметить, что в отдельных моделях значения коэффициентов и их значимость (p-value) различаются для белого и красного вина.
Например:

- total sulfur dioxide имеет значительное влияние на качество красного вина (p-value = 0.000), но не значимо для белого.
- fixed acidity значим для белого вина (p-value = 0.000), но не оказывает значительного влияния на качество красного вина (p-value = 0,722).

Это подтверждает наличие существенного различия во взаимосвязях признаков с красными и белыми винами.

Выводы по гипотезам:

Во всех трех моделях коэффициент при переменной alcohol статистически значим (p-value < 0.05) и положителен:

- для красного вина: β=0.3469
- для белого вина: β=0.2425
- в общей модели: β=0.2747

Это подтверждает, что алкоголь действительно оказывает положительное влияние на оценку качества вина, особенно красного. Значит Гипотеза 1 подтверждается.

Коэффициент при переменной volatile acidity статистически значим (p-value < 0.05) и отрицателен во всех моделях:

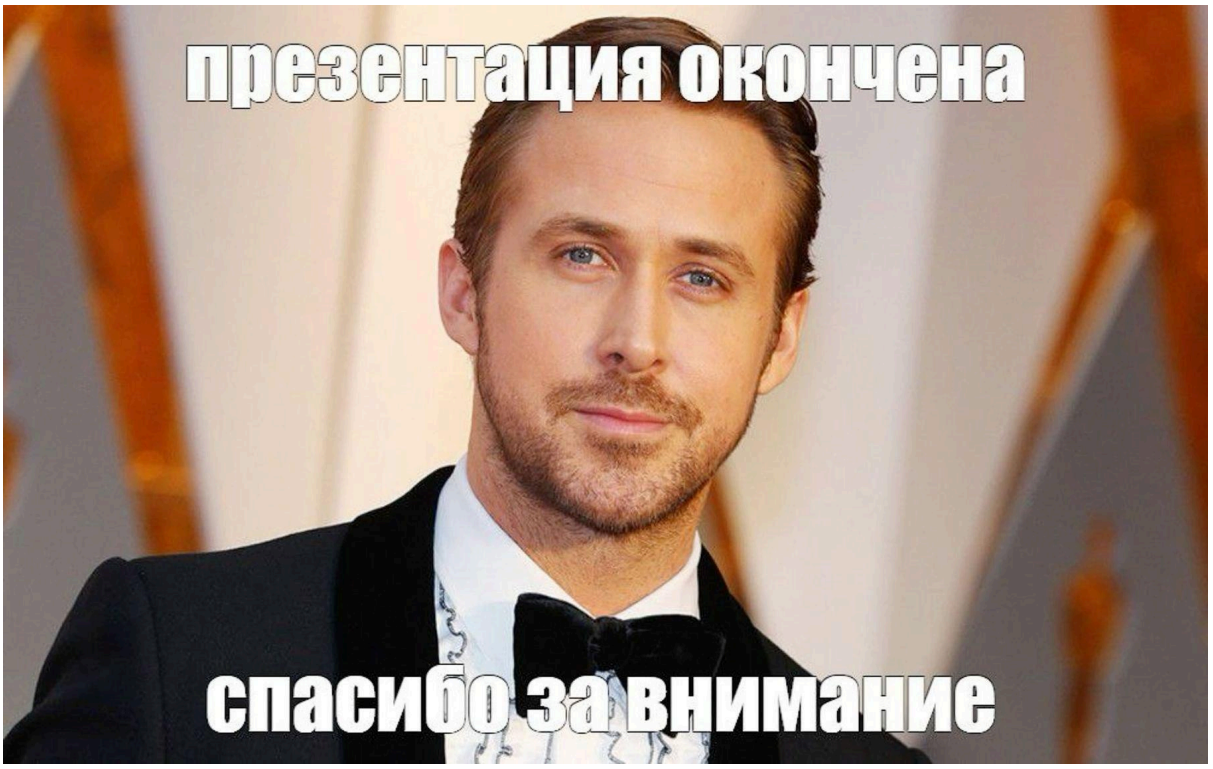
- для красного вина: β=−0.1895
- для белого вина: β=−0.2698
- в общей модели: β=−0.2264

Это подтверждает, что высокая летучая кислотность снижает качество вина, причем на качество белого вина она влияет сильнее. Значит Гипотеза 2 тоже подтверждается.

Коэффициент при переменной residual sugar положителен во всех моделях, но не везде значим:

- для красного вина: β=0.0354, незначимый(p-value = 0.660)
- для белого вина: β=0.3292 , значимый (p-value < 0.05)
- в общей модели: β=0.2475, значимый (p-value < 0.05)

На качество белого вина остаточный сахар влияет положительно, добавляя баланс и улучшая вкусовой профиль. На качество красных вин остаточный сахар особо не влияет (так как коэф незначим). Значит Гипотеза 3 частично подтверждается.



Используемые источники

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Modeling wine preferences by data mining from physicochemical properties*. Decision Support Systems, 47(4), 547-553. https://repositorium.sdum.uminho.pt/bitstream/1822/10029/1/wine5.pdf

[2] Wang, H., Peng, H., Zhang, L., & Zhang, Z. (2015). Chemometric methods for comprehensive review of wine quality based on physicochemical properties and volatile compounds. *Molecules, 20*(5), 8453–8483. https://doi.org/10.3390/molecules20058453

[3] Zamuz, S., Rodríguez-Bermúdez, R., Rocchetti, G., Barba, F. J., & Lorenzo, J. M. (2023). Effect of winemaking process on phenolic composition and antioxidant capacity of wines: A review. *Beverages, 9*(2), Article 41. https://doi.org/10.3390/beverages9020041

[4] Agyemang, P. (2016). *Predicting wine quality using machine learning techniques*. [Master's thesis, Youngstown State University]. YSU Digital Repository. https://digital.maag.ysu.edu/xmlui/bitstream/handle/1989/10638/Agyemang%20Perpetual%20APPROVED.pdf?sequence=3

[5] *Wine Quality Dataset*. Kaggle. https://www.kaggle.com/datasets/rajyellow46/wine-quality/data