

Университет ИТМО

Практическая работа №2
по дисциплине «Визуализация и моделирование»

Автор: Литвиненко Даниил Дмитриевич

Поток: ВИМ 1.2

Группа: К3221

Факультет: ИКТ

Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Выбранный датасет содержит информацию об основных характеристиках вин и оценки их качества.

Название столбца	Данные	Тип данных
fixed acidity	содержание нелетучих кислот	float
volatile acidity	содержание летучих кислот	float
citric acid	содержание лимонной кислоты	float
residual sugar	количество сахара, оставшегося после остановки брожения	float
chlorides	количество соли в вине	float
free sulfur dioxide	свободная форма SO ₂ находящаяся в равновесии между молекулярным SO ₂ (в виде растворенного газа) и бисульфит-ионом	float
total sulfur dioxide	количество свободных и связанных форм S ₀₂	float
density	Относительная плотность между плотностью воды	float
pH	Кислотность вина	float
sulphates	винная добавка, которая может способствовать повышению уровня газообразного диоксида серы (S ₀₂)	float
alcohol	процентное содержание алкоголя в вине	float
quality	Конечная оценка (на основе сенсорных данных, оценка от 0 до 10)	float

В датасете все данные представлены в числовом формате, поэтому форматирование не требуется. Однако датасет включает два файла с данными о красном и белом вине, поэтому они были объединены в один датафрейм с дополнительным полем содержащим тип вина.

Первое на что можно хочется узнать, это распределение оценок.

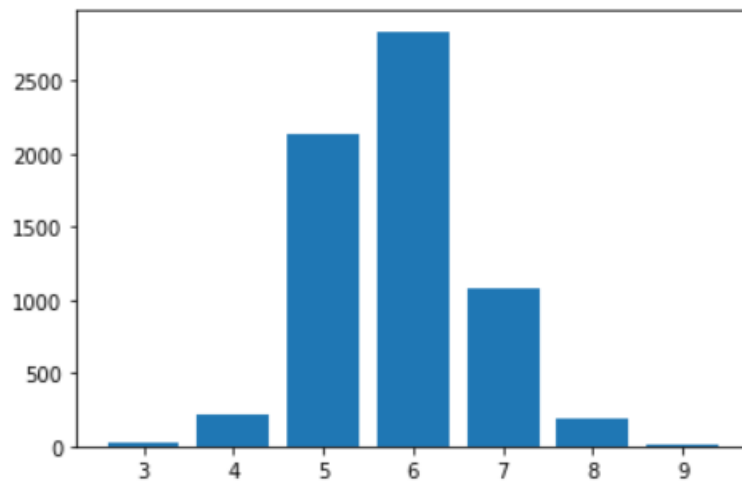


Рис. 1: Распределение оценок

Как видно из гистограммы, самые частые оценки - 5, 6. Также мы понимаем что в датасете ни одно вино не имеет оценку 10 или ниже 3.

Рассмотрим отдельно на красное и белое вино.

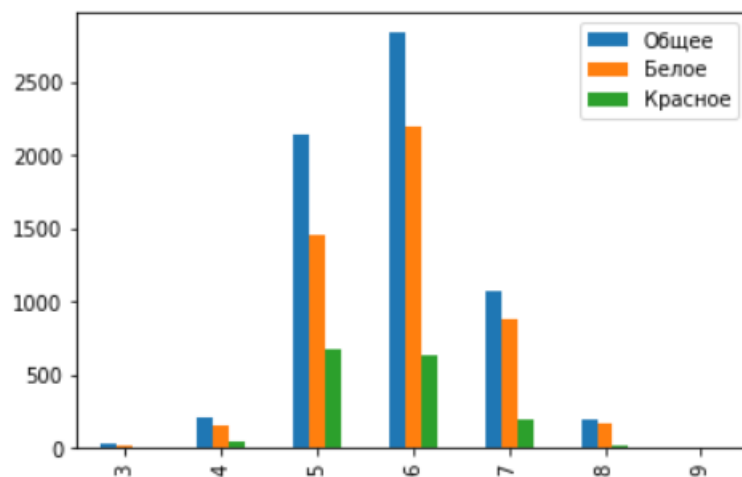


Рис. 2: Распределение оценок для красного и белого вина.

Данная гистограмма не даёт полного представления о распределении оценки. Поэтому имеет значение рассмотреть распределение оценок в процентах.

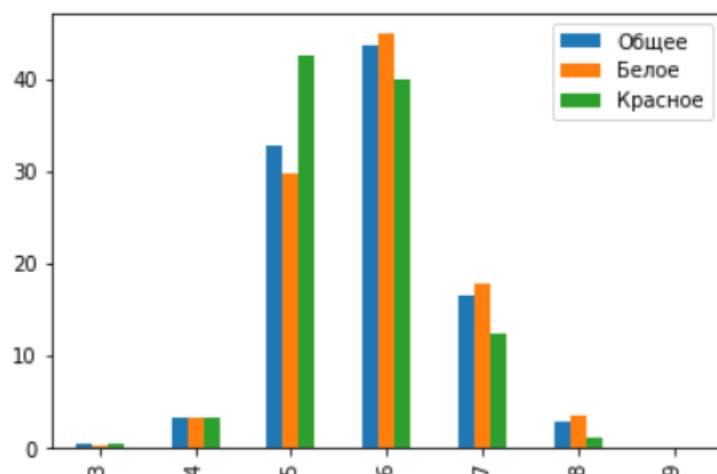


Рис. 3: Распределение оценок для красного и белого вина в процентах.

Как видно, Красное вино имеет меньше высоких оценок, чем белое, из чего можно сделать вывод, что как правило, белое вино оценивается лучше чем красное.

В нашем датасете имеются данные по содержанию различных кислот в вине, также имеется показатель кислотности pH. Посмотрим, какая кислота вносит наибольший вклад в кислотность. Для этого, для каждой кислоты построим точечный график, где по горизонтале уровень pH, а по вертикали содержание кислоты.

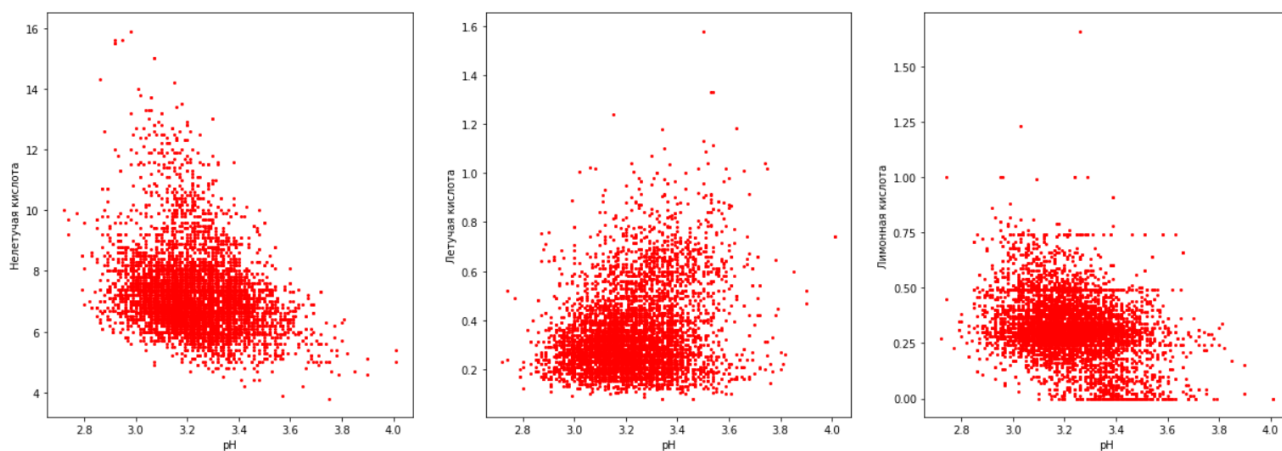


Рис. 4: Содержание кислот и pH

Лучше всего видна корреляция для нелетучей кислоты.

Считается, что одним из показателей качества вина является содержание летучих кислот, так как они образуются в большом количестве при проведении неправильной процедуры брожения, или при использовании винограда, содержащего повышенное количество болезнетворных бактерий. Так как оценки качества у нас нет, то проанализируем оценку вин в зависимости от содержания летучей кислоты. Для этого воспользуемся простым методом. Найдём средний показатель летучей кислоты и разделим датасет на две части: с содержанием летучей кислоты выше среднего и ниже. Найдём среднюю оценку для обеих групп.

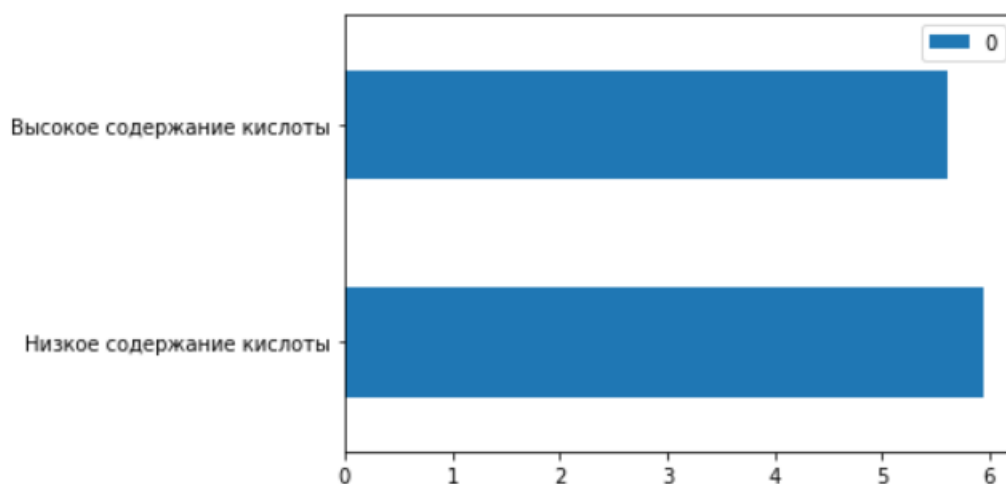


Рис. 5: Средняя оценка для разных уровней летучей кислоты

Выяснилось что количество летучей кислоты не влияет на оценку вина.

Проверим ещё одну гипотезу. Очень кислые вина, как и чрезмерно пресные, должны оцениваться ниже, чем вина с нормальным уровнем кислотности. Для этого разобьём все вина на 20 частей с разным уровнем кислотности, от самых пресных, до самых кислотных. Затем посчитаем среднюю оценку для каждой группы.

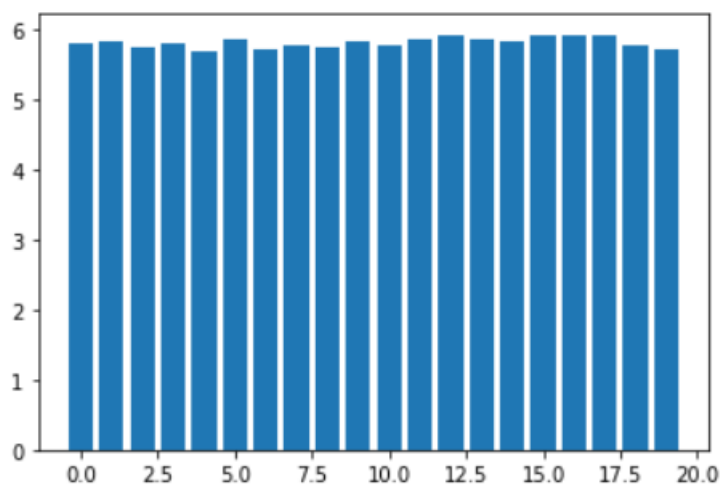


Рис. 6: Средняя оценка для разных уровней кислотности

Выяснилось, что уровень кислотности никак не влияет на оценку вина.

Из базового курса биологии и химии, мы знаем, что при брожении часть сахара преобразуется в этиловый спирт. Спирт напрямую влияет на уровень алкоголя в вине. Также в датасете есть количество сахара, оставшегося после остановки брожения. Посмотри как коррелируют эти две величины. Ожидается, чем меньше сахара - тем меньше содержание алкоголя. Для проверки гипотезы построим точечную диаграмму.

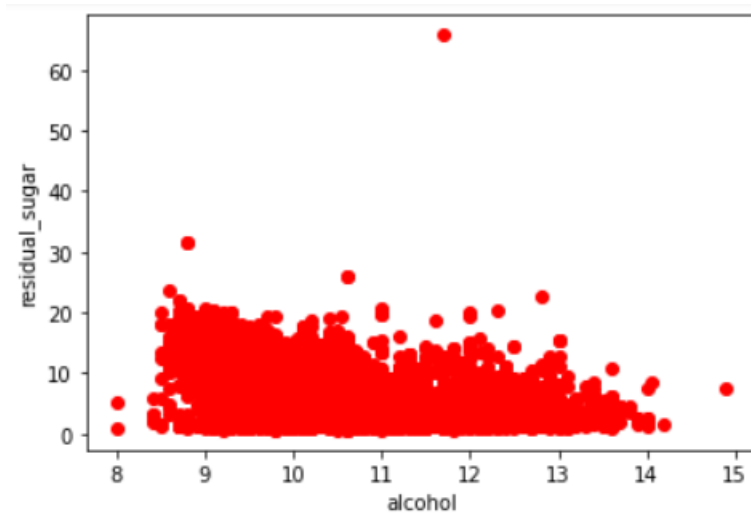


Рис. 7: содержание алкоголя и сахара после брожения

На графике мы видим "горку". Мы можем предположить, что "гребень горки" сладкие вина, так как они имеют наибольшее количество сахара среди прочих, "подножье горки" сухие, так как уровень сахара у них 0. А между - полусухие.

Неотходя от темы алкоголя, выясним, каких вин больше - слабоалкогольных или сильноалкогольных. Разделим вина на 4 категории по содержанию алкоголя (равномерно). Проанализируем количество вин в каждой группе.

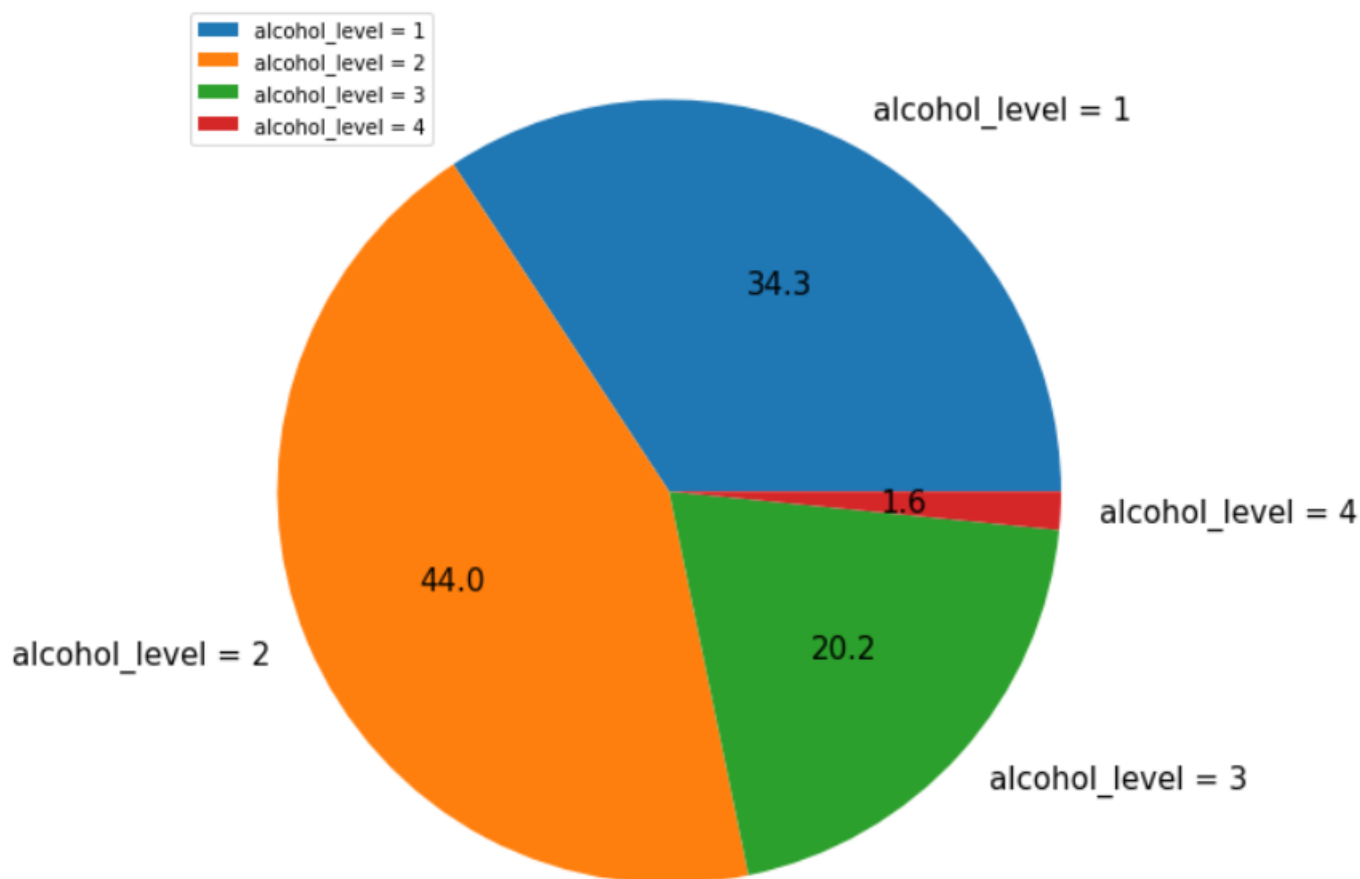


Рис. 8: Количество вин в категориях алкогольности

Слабоалкогольных вин больше чем крепких.

Проанализируем теперь сульфатное содержание вин.

Во многие вина добавляют добавки для образования диоксида сера, для более выраженного вкуса и небольшого количество газа.

В датасете есть показатели свободного SO₂ и общего SO₂. Посмотрим на какой диоксид больше всего влияет добавка. Построим два точечных графика.

Из них видно, что четкой зависимости нет. Следовательно добавка не вносит большой вклад в количество свободного и связанного SO₂. Однако, видна Общая "горка" поэтому имеет смысл проверить корреляцию между свободным общим количеством SO₂.

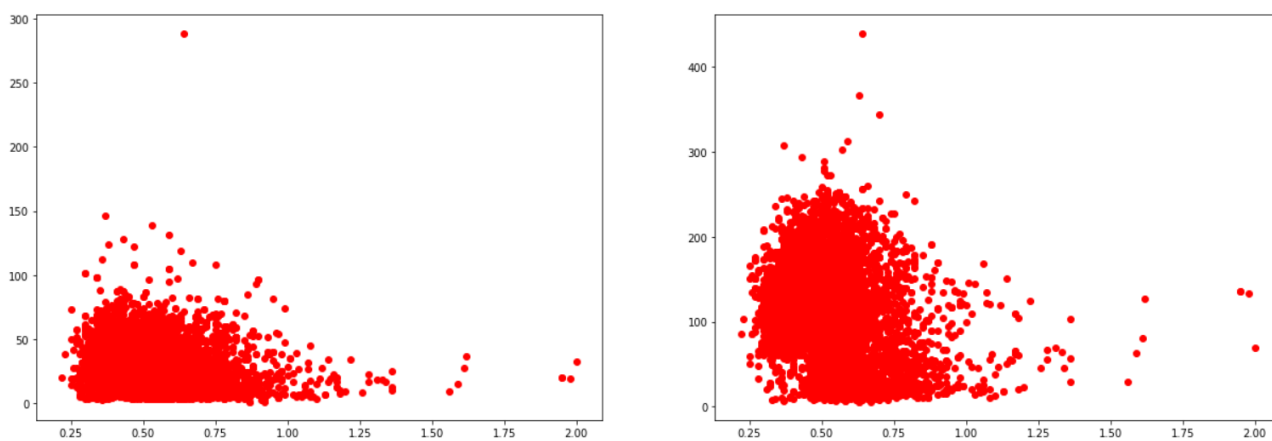


Рис. 9: содержание SO₂ в зависимости от количества виинной добавки

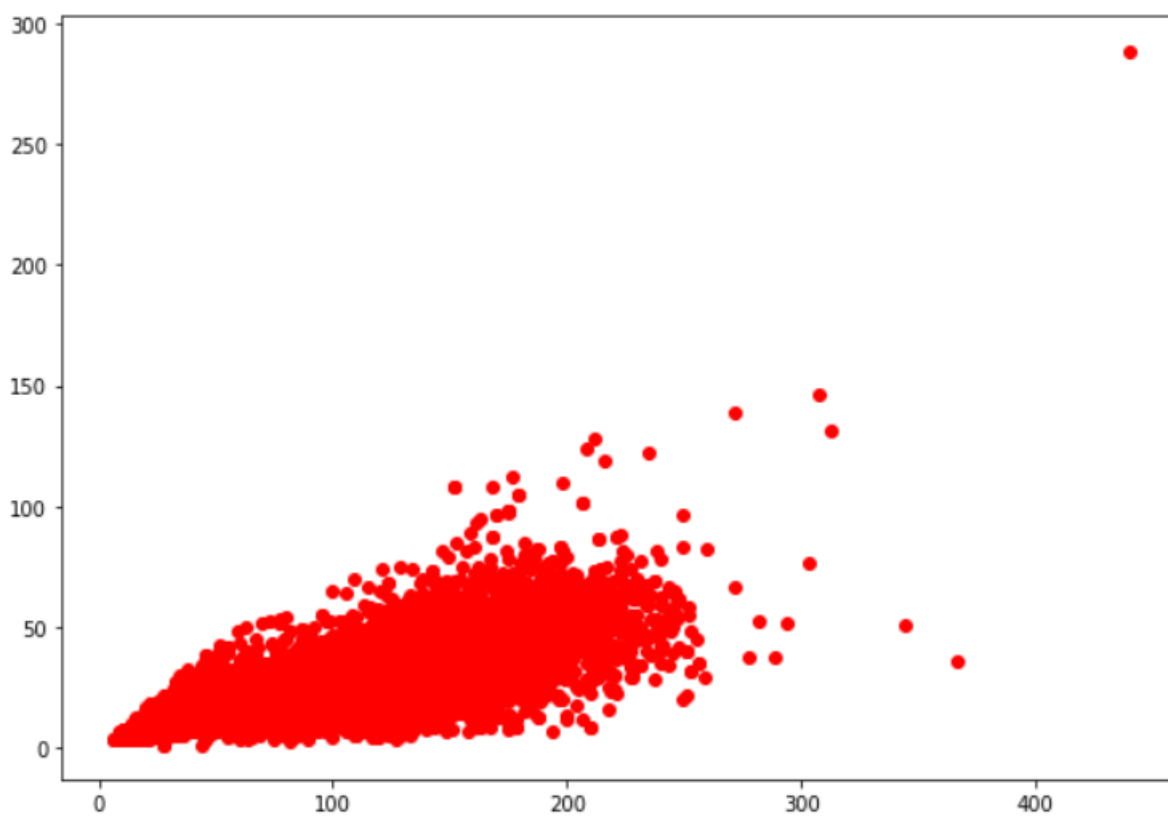


Рис. 10: Зависимость между связанным SO₂ и Общим количеством SO₂