**НИЯУ “МИФИ”**

**Контрольное задание №1**

Группа: М19-117

Студент: Кайгородов Александр

Преподаватель: Попов И.В.

Москва, 2020

**ЗАДАНИЕ:**

1. Собрать данные для парной регрессии (линейной или нелинейной).

2. Провести исследования аналогично, как в лекции.

3. Оформить работу в Word c шапкой номером группы и ФИО.

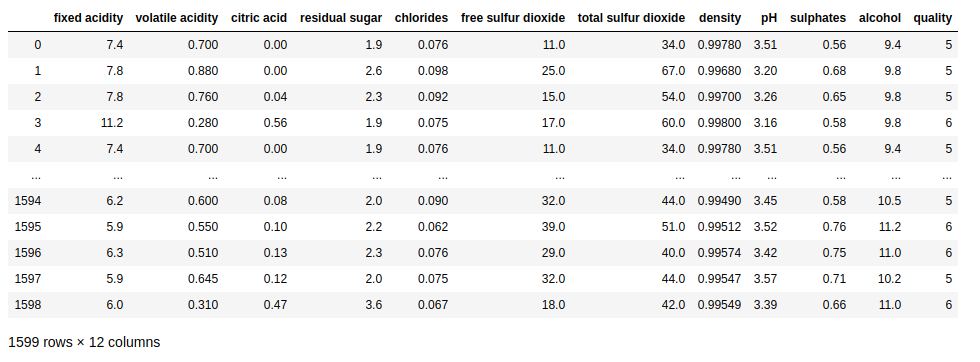
**РЕШЕНИЕ:**

Взятые данные были предоставлены в качестве исходных данных для одного из существующих соревнований на сайте Kagge

**Context**

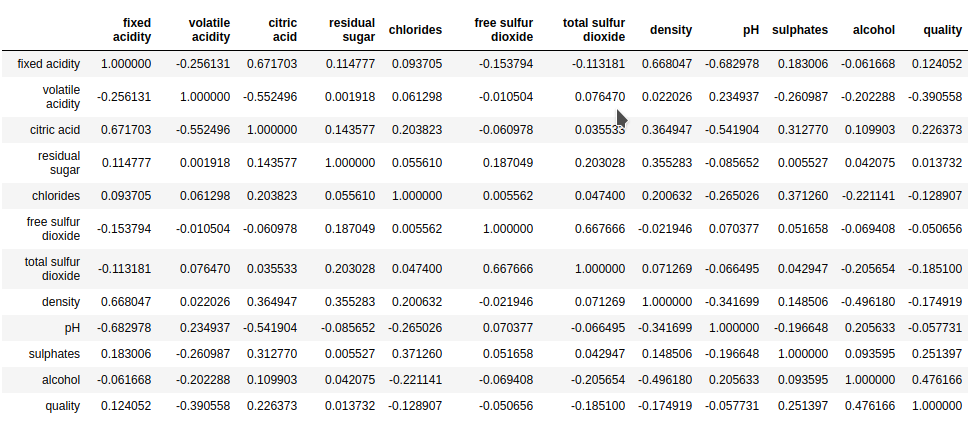
The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

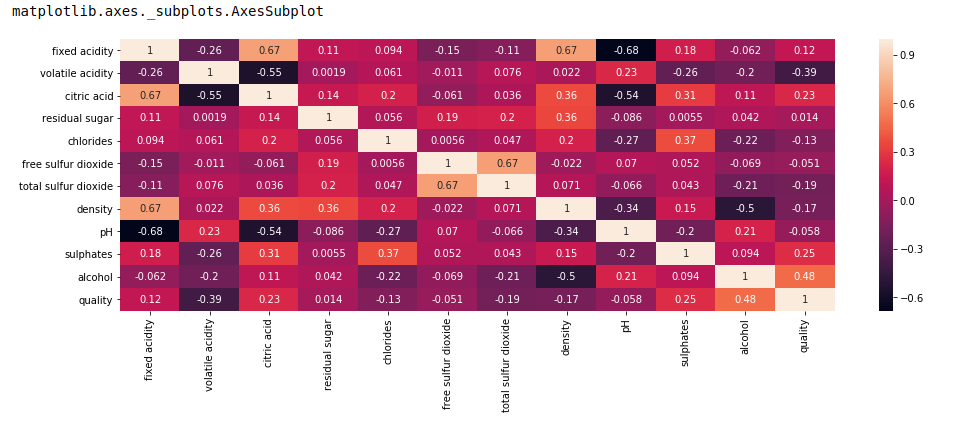


**Поиск любопытных зависимостей (Mining)**

Можно построить матрицу корреляций для выявления наиболее коррелирующих признаков. Определив наиболее коррелирующие величины можно в принципе говорить о поиске, возможно "важных", зависимостей.

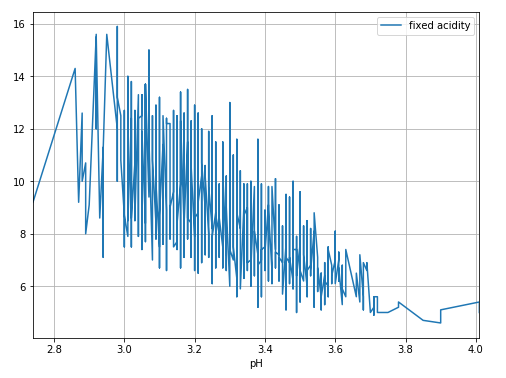


Так же можно построить тепловую карту (heatMap) для имеющихся признаков. Иногда такая визуальная составляющая может упростить поиск зависимостей для человека.



Наиболее коррелирующее значение принадлежит показателям кислотности и рН вина. Как видно по графику, изменение одной величины обратно пропорционально изменению другой. Это в принципе и так ясно исходя из значения коэффициента корреляции.

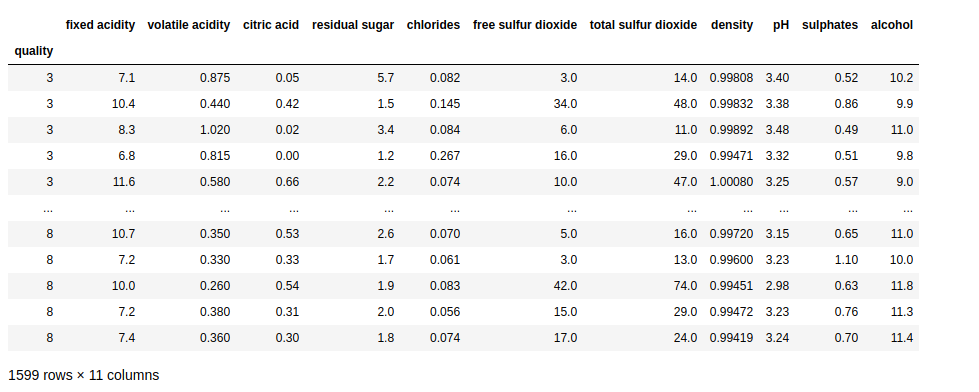
**r(pH, fixed acidity) = -0.682978**



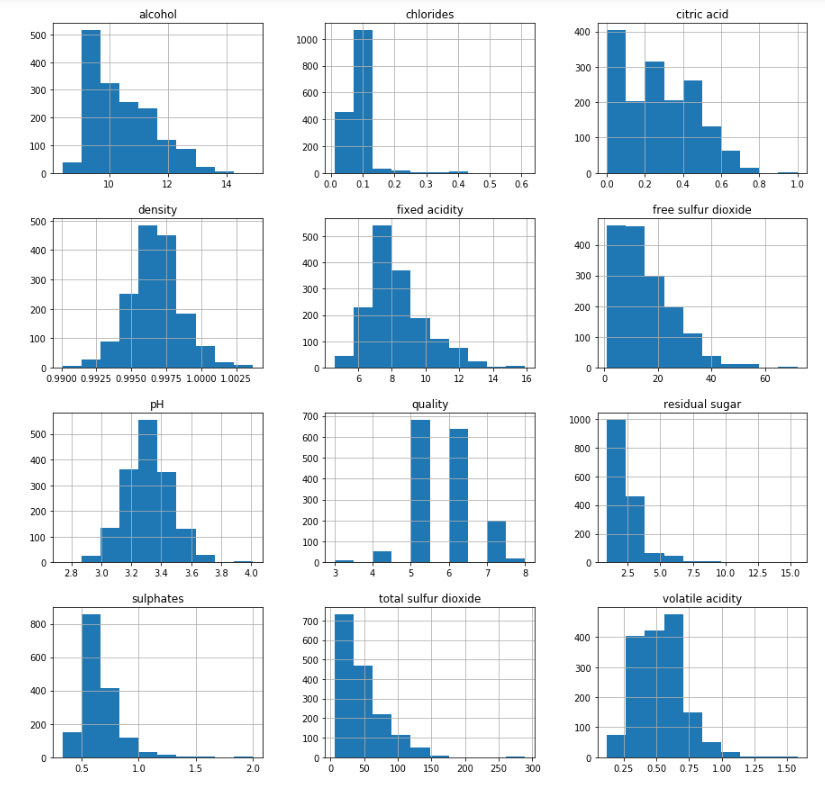
Но зависимость может оказаться случайной, продолжим исследование

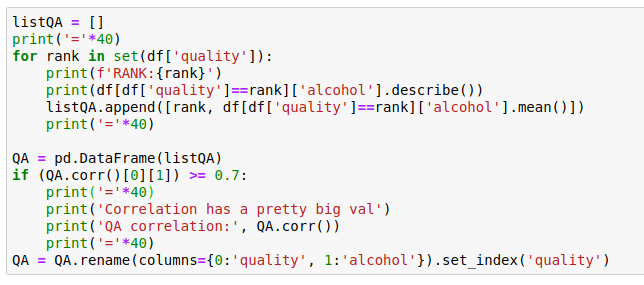
Для анализа может оказаться весьма интересно узнать зависимость поставленного качества от величины какого либо параметра

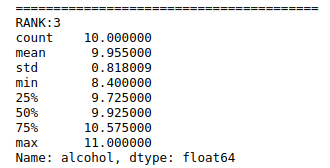
Проведем сравнение количественных признаков относительно атрибутивного признака представленного в виде значения "КАЧЕСТВО".



Наблюдает гистограммы распределения числа значений величин в данных

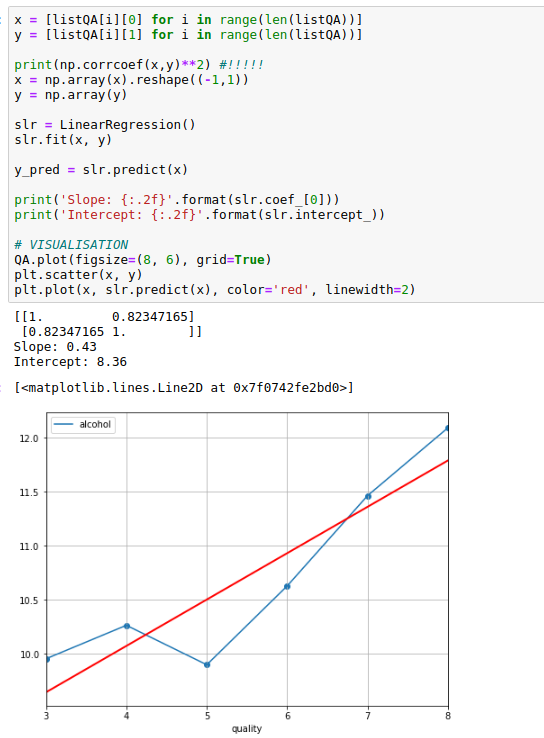


Проведем анализ распределения крепкости для каждого уровня качества: 



Аналогично, в прикрепленном файле PDF приведены результаты расчетов по остальным баллам качества.

# **Регрессионный анализ**

На первом этапе линейного регрессионного анализа необходимо получить непосредственно саму регрессионную прямую. Воспользуемся библиотекой scikit.learn построив модель линейной регрессии и получив значения коэффициентов а и b

**Уравнение прямой регрессии:**

𝑎𝑙𝑐𝑜ℎ𝑜𝑙=0.43∗𝑞𝑢𝑎𝑙𝑖𝑡𝑦+8.36

Вывод: Более крепкое вино зачастую получается большую оценку качества чем аналоги с меньшей крепкостью.

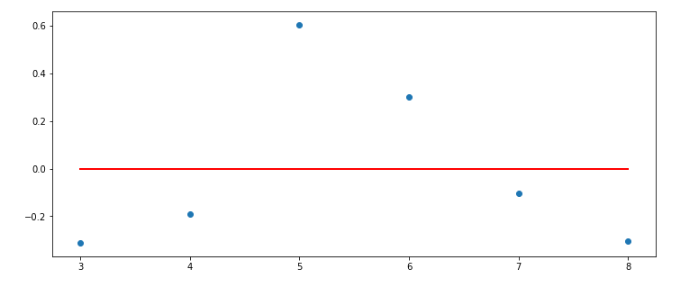
*PS: слегка не логично выбирать в качестве варьируемого параметра именно качество и получать по нему крепкость, но сути дела это не меняет, зависимость одного параметра от другого получена.*

Расположение отклонений истинных значений от предсказанных. Судя по данному графику можно положить что данная зависимость может подвергаться линейному регрессионному анализу

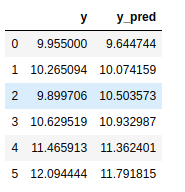
Для более наглядного представления изобразим данный факт на графике отклонений

СИНИЕ ТОЧКИ - истинные значения

КРАСНАЯ ЛИНИЯ - предсказываемые линейной регрессией значения



**Определение коэффициента детерминации**

****

r^2 (coefficient of determination): 0.823

Значение получает верное значение.   
Коэфф корреляции уже был вычислен и равен r = 0.907453

Стандартные ошибки

Se = 0.115

Ser\_a = 0.0576

Ser\_b = 0.0198

Проверка полученных коэффициентов на значимость

На основании имеющихся наблюдений построена зависимость оценки качества алкоголя от его крепкости. В действительности зависимым параметром выбрано качество, но в целом сути дела это не меняет и зависимость описывается следующим уравнением прямой.  
  
Y’ = 0.43⋅𝑥+8.36

Ser(a) = 0.0576

Ser(b) = 0.0198

Проверим гипотезы о значимости коэффициентов на уровне значимости 5%:

* Формулировка гипотез:  
  H0: a = 0; H1: a != 0;  
  H0: b = 0; H1: b != 0;
* Критические значения для распределения Стьюдента с n-2 = 4 степенями свободы равны ta/2 = 2.77645. Следовательно, критическая область |t| > 2.77645 (Гипотеза двухсторонняя)
* Определяем расчетные значения критерия t(b):

tb = 422.97823464476386

ta = 7.459219722375194

* Принятие решения  
  Расчетные значения критерия для обоих коэффициентов превышают критическое значение:  
    
  t(b)> ta/2, 423 > 2.776;  
  t(a)> ta/2, 7.5 > 2,776;  
  -> основные гипотезы отклоняются
* Вывод  
  Оба коэффициента регрессионной модели значимы (не случайные) <- отличаются от нуля.

# Заключение

частичное исследование представленных данных показало что выбор более крепкого сорта вина увеличивает вероятность того что это вино окажется лучшего качества. Регрессионный анализ показывает о весьма сильной корреляции данных величин между собой, следовательно данный фактор может оказаться весьма весомым при выборе спиртного напитка.

Полученные значения, при проверке, не оказались случайными и могут считаться действительными.