Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра математики

Курсова робота

освітній ступінь – бакалавр

на тему: «Дослідження методу основних компонент для зменшення розмірності»

Керівник курсової роботи

Старший викладач,

кандидат фізико-математичних наук

Дрінь Світлана Сергіївна

_____ (підпис)

Виконала студентка 3-го року навчання,

напрямку підготовки

Прикладна математика

Чернявська Карина Сергіївна

Рецензент _____

Кваліфікаційна робота захищена

з оцінкою: _____

Секретар ЕК_____

«____» _____2022р.

Київ – 2022

Календарний план виконання роботи:

№ п/п	Назва етапу дипломного проекту (роботи)	Термін виконання етапу	Примітка
1.	Отримання завдання на курсову роботу.	11.10.2021	
2.	Огляд технічної літератури за темою роботи.	20.11.2021	
3.	Виконати аналіз методу	28.11.2021	
3.	Розробка моделі	12.12.2021	
4.	Розробка алгоритму	24.01.2022	
5.	Програмування розробленого алгоритму	15.02.2022	
6.	Написання пояснювальної роботи.	31.05.2022	
7.	Створення слайдів для доповіді та написання доповіді.	04.06.2022	
9.	Аналіз отриманих результатів з керівником, написання доповіді та попередній захист курсової роботи.	06.06.2022	
10.	Корегування роботи за результатами попереднього захисту.	05.06.2022	
11.	Остаточне оформлення пояснювальної роботи та слайдів.	06.06.2022	

Студент __Чернявська К.С._____

Керівник __Дрінь С.С._____

“_____”_____

Зміст

Вступ..... 4
Постановка задачі 6
Розділ 1. Історичні відомості методу головних компонент..... 7
Розділ 2. Основні теоретичні відомості. 8
Метод головних компонент. 8
Розділ 3. Алгебраїчний приклад методу головних компонент 10
Висновок 17
Перелік використаних джерел 18
Додаток..... 19

Вступ

З даними бувають дві проблеми: або їх дуже мало, або їх дуже багато. Сьогодні ми поговоримо про другу проблему.

З числа методів, що дозволяють узагальнювати значення елементарних факторів, метод головних компонент виділяється простою логічною конструкцією; на його прикладі стають зрозумілими загальна ідея і цільові настанови численних методів факторного аналізу. Для виявлення найбільш значущих факторів і їх структури, найбільш виправданим є метод головних компонент.

Сутність методу полягає в заміні корельованих компонентів некоррельованими факторами. Іншою важливою характеристикою методу є можливість виокремлення найбільш інформативних головних компонент і виключення інших з аналізу, що спрощує інтерпретацію результатів.

Датасет складається з m вхідних факторів. Метод головних компонент дає можливість за m вхідними факторами виділити m головних компонент. Досить часто головні компоненти називаються узагальненими факторами. Простір головних компонентів є ортогональним. Математична модель головних компонент базується на логічному допущенні, що значення безлічі взаємозалежних факторів породжують деякий загальний результат.

Метод головних компонент використовується для вивчення взаємозв’язків між досліджуваними показниками. За його допомогою можна виявляти приховані зв’язки між факторами, які відповідають за наявність лінійних статистичних зв’язків (кореляцій) між ними. Крім того, визначення більш впливових за умов проведення досліджень факторів серед первинно обраних показників, а також виявлення статистичного зв’язку визначають обґрунтованість висновків щодо ефективності тих чи інших впливів на досліджувану систему.

Метод головних компонент розв'язує такі завдання:

1. Виявлення прихованих, об'єктивно існуючих закономірностей у зміні явищ.

2. Явище, яке ми вивчаємо на первинному етапі характеризується великою кількістю факторів, які можна описати лінійною комбінацією меншої кількості факторів. Число головних компонент, виділених в процесі дослідження, буде у компактній формі вміщувати (компактна форма – лінійна комбінація факторів) більше інформації, ніж початковому етапі.

3. Виявлення факторів, найбільш тісно пов'язаних з головною компонентою. Інакше кажучи, вивчення стохастичного зв'язку між ними.

4. Прогнозування явищ на підставі рівняння регресії, яке базується на основі головних компонент.

Постановка задачі

«Словарь Вильяма Шекспира, по подсчёту исследователей, составляет 12000 слов. Словарь негра из людоедского племени «Мумбо-Юмбо» составляет 300 слов. Эллочка Щукина легко и свободно обходилась тридцатью.» - за цитатою І. Ілля та Е. Петров. 12 стільців

Розділ 1. Історичні відомості методу головних компонент

Вперше він був розроблений в 1901 р. англійським статистом К.Пірсоном . Потім знайшов свій розвиток у працях Г.Хотеллінга, Г.Хармана, С. Рао, П.Андрукевича, С.Айвазяна і інших авторів.

Доповнений і розширений Гарольдом Готелінґом в 1933 р. Застосовується в багатьох галузях, зокрема,
в економетриці, біоінформатиці, обробці зображень, для стиснення даних, у суспільних науках.

До недавнього часу метод головних компонент вважали різновидом факторного аналізу. Нині його відносять до групи самостійних статистико - математичних методів багатомірного аналізу.

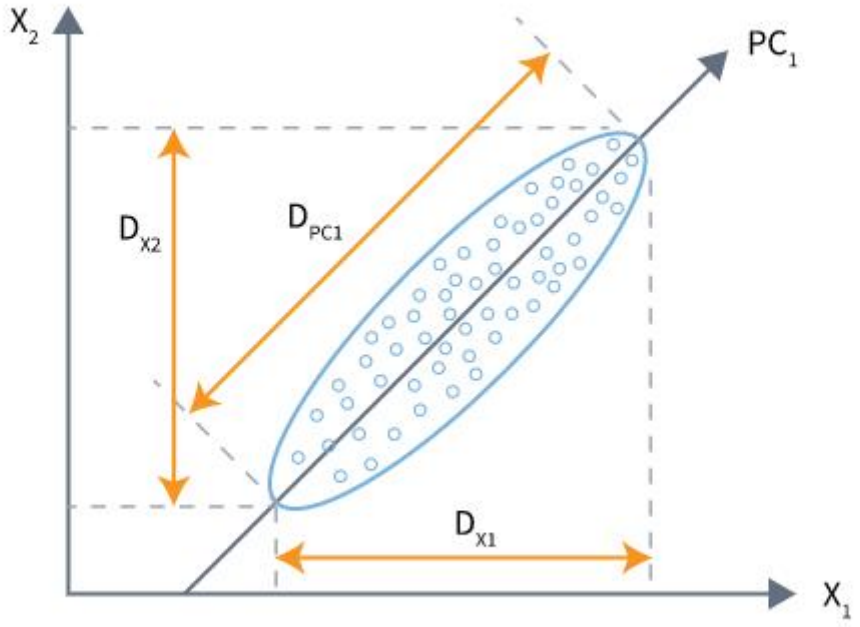
Розділ 2. Основні теоретичні відомості.

Метод головних компонент.

Лінійне перетворення A евклідового простору V називається ортогональним, якщо воно зберігає скалярний добуток векторів, тобто $\forall\ x,\ y \in V\ (Ax,\ Ay) =\ (x,\ y)$. Це означає , що ортогональне перетворення зберігає довжини векторів та кути між ними.

З математичної точки зору метод головних компонент являє собою ортогональне лінійне перетворення, яке відображає дані з вихідного простору у новому просторі меньшої розмірності.

При цьому перша вісь нової системи координат будується таким чином, щоб дисперсія даних була мінімальною. Друга вісь будується ортогонально до першої так, щоб її дисперсія вздовж першої, була максимальною можливою з існуючих і т.д. Перша вісь називається першою головною компонентою, друга — другою і т.д.



Мал. 2.1

На малюнку 2.1 показано зниження розмірів вихідного 2-мірного простору (X_1, X_2) за допомогою методу головних компонентів до 1-мірного. Перша головна компонента PC1 орієнтована вздовж напряму найбільшої витянутості еліпсоїда розсіяння точок об'єктів вихідного набору даних в просторі визнань, з нею пов’язана найбільша дисперсія.

На малюнку 1 також неважко побачити, що проекція дисперсії даних на вісь першої частини компонента D_{pc} більше, ніж її проекції на вихідні осі D_{x1} та D_{x2} , але менше їх суми. Тобто, з головними компонентами вираження усієї дисперсії даних не вдалося. Тому будуємо другу основну компоненту.

Отже, сенс методу у тому, що з кожного вхідного фактора визначається певна частка дисперсії вихідного набору даних. У свою чергу дисперсія, що є мірою мінливості даних, може відображати рівень їх інформативності.

Розділ 3. Алгебраїчний приклад методу головних компонент

Я вважаю, що математична інтуїція дозволяє нам з'єднувати шматочки головоломки в довгостроковій перспективі. Розглянемо приклад покрокового розрахунку PCA на прикладі.

Скажімо, у нас є набір даних про результати іспитів 5 учнів за їхні різні спортивні уроки, і ми спробуємо зменшити розмір цього набору даних.

	Tennis	Volleyball	Running	Label
Student - 1	6	4	4	0
Student - 2	8	3	3	1
Student - 3	5	8	4	1
Student - 4	4	5	10	0
Student - 5	2	8	5	0

Табл. 3.1

1 крок - Візьміть особливості набору даних, не враховуючи Label.

PCA працює над зниженням розмірності (кількості вимірів) даних, які в даному випадку є заняттями спорту. Тому ми не повинні розглядати стовпець мітки та унікальні ідентифікатори студента. Цей процес пов'язаний лише з особливостями набору даних. Після виконання всіх кроків результатом стануть нові бали кожного учня у вибраних параметрах PCA.

$$A = \begin{bmatrix} 6 & 4 & 4 \\ 8 & 3 & 3 \\ 5 & 8 & 4 \\ 4 & 5 & 10 \\ 2 & 8 & 5 \end{bmatrix}$$

2 крок - Нормалізація набору даних

Новий простір – нова лінійна комбінація

PCA – це практика зміни напрямку компонентів на напрямки максимальної дисперсії.

Нормалізація проводиться в два етапи:

Перш за все, віднімаємо відповідні середні значення змінної кожного набору даних. Це створює набір даних, середнє значення якого дорівнює нулю.

Ви можете побачити середнє значення кожного стовпця, як показано нижче:

$$\bar{A} = [5.0,5.6,5.2]$$

По-друге, поділ усіх даних вказує на стандартне відхилення.

3 крок - Розрахувати коваріаційну матрицю

Різниця між коваріацією та дисперсією. Дисперсія вимірює коливання однієї випадкової величини (наприклад, зріст людини в популяції), тоді як коваріація — це вимірювання того, наскільки дві випадкові величини змінюються разом (наприклад, зріст людини та вага людини в популяції).

Нормалізація точок даних пов'язана з коваріаційною матрицею. Оскільки у формулі коваріації середнє значення кожного стовпця (фактора) потрібно відняти з кожного елемента даних кожного атрибута. Якщо застосувати нормалізацію, середнє значення буде нульовим.

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y})$$

Розрахунок коваріації має відбуватися в комбінації кожних двох факторів.

$$\text{Covariance Matrix } = \begin{bmatrix} \mathrm{cov(X,X)} & \mathrm{cov(X,Y)} & \mathrm{cov(X,Z)} \\ \mathrm{cov(Y,X)} & \mathrm{cov(Y,Y)} & \mathrm{cov(Y,Z)} \\ \mathrm{cov(Z,X)} & \mathrm{cov(Z,Y)} & \mathrm{cov(Z,Z)} \end{bmatrix}$$

Якщо ми скажемо, що маємо вимір d(кількість факторів), коваріаційна матриця буде у розмірах $[d \times d]$.

- Це квадратна матриця
- Це симетрична матриця відносно діагоналі, тому

$$cov(X,Y) = cov(Y,X)$$

$cov(X,X)$ вказує на дисперсію фактора X.

Чим більше значення діагонального елемента в матриці, тим більша дисперсія. Тому діагональна частина матриці вказує на дисперсію кожного фактора. Усі інші елементи в матриці вказують на коваріацію двох різних змінних.

$$\text{Covariance Matrix} = \begin{bmatrix} \text{cov}(X,X) & \text{cov}(X,Y) & \text{cov}(X,Z) \\ \text{cov}(Y,X) & \text{cov}(Y,Y) & \text{cov}(Y,Z) \\ \text{cov}(Z,X) & \text{cov}(Z,Y) & \text{cov}(Z,Z) \end{bmatrix}$$

Повертаючись до нашого прикладу, де

$$A = \begin{bmatrix} 6 & 4 & 4 \\ 8 & 3 & 3 \\ 5 & 8 & 4 \\ 4 & 5 & 10 \\ 2 & 8 & 5 \end{bmatrix} \qquad \bar{A} = [5.0, 5.6, 5.2]$$

Розрахуємо коваріацію:

Cov(X,Y) = [(6–5)*(4–5.6) + (8–5)*(3–5.6) + (5–5)*(8–5.6) + (4–5)*(5–5.6) + (2–5)*(8–5.6)] / (5–1)= -4

$$\text{Covariance Matrix} = \begin{bmatrix} 5 & -4 & -3 \\ -4 & 5.3 & 0.35 \\ -3 & 0.35 & 7.7 \end{bmatrix}$$

Коваріаційна матриця вказує на :

- Найбільше коливання (дисперсія) – у бігу (+7,7), а найменше – у тенісних (+5,0) тренуваннях.
- Коваріація між волейболом і бігом є позитивною (0,35), це означає, що результати мають позитивний зв’язок, хоча і не дуже сильний. Наприклад, якщо результати в тенісі зростають, результати в бігу також зростають, і навпаки.
- Коваріація між тенісом і волейболом від'ємна (-4), це означає, що фактори мають негативний зв’язок. Тобто, результати в тенісі зростають, результати у волейболі падають, і навпаки.
- Якщо коваріація будь-яких двох факторів дорівнювала нулю, це означає, що немає чіткого зв’язку.

4 крок - Обчислимо власні значення

Власне значення — це скаляр, який використовується для перетворення власного вектора.

$$A \cdot v = \lambda \cdot v \tag{1}$$

A — квадратна матриця, v — власний вектор,

λ — власне значення, пов'язане із власним вектором матриці A.

Коли ви перенесете праву частину рівняння вліво, вам потрібно розв'язати наведене нижче рівняння, щоб отримати власне значення λ.

$$\det(A - \lambda \cdot E) = 0$$

$$\det\left(\begin{pmatrix} 5 & -4 & -3 \\ -4 & 5.3 & 0.35 \\ -3 & 0.35 & 7.7 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) = 0$$

$$\det\left(\begin{pmatrix} 5 - \lambda & -4 & -3 \\ -4 & 5.3 - \lambda & 0.35 \\ -3 & 0.35 & 7.7 - \lambda \end{pmatrix}\right) = 0$$

Тепер нам потрібно отримати значення λ розв'язавши наведене вище рівняння.

Рівняння буде мати вигляд:

$$-\lambda^3 + 18\lambda^2 - 80\lambda + 40 = 0$$

Розв'язавши це рівняння, ви отримаєте:

$$\lambda \approx 0.58 \quad \lambda \approx 11.0 \quad \lambda \approx 6.4$$

5 крок - Обчисліть власні вектори

Власні вектори - це вектори, помножені на власне значення в лінійних перетвореннях матриці. Власні значення - це константи, які множать власні вектори в лінійних перетвореннях матриці. Власні вектори перекладають інформацію з вихідної матриці в множення значень і константу. Власні значення - це ця константа, яка множить власні вектори та бере участь у лінійному перетворенні вихідної матриці.

Для знаходження власних значень використовуємо рівняння (1).

Для кожного значення λ знаходимо відповідний власний вектор.

$$A \cdot v = \lambda \cdot v$$

$$(A - \lambda \cdot E) \cdot v = 0$$

$(A - \lambda \cdot E) = B$, де матриця B дорівнює A мінус матриця власних значень.

$$B \cdot v = 0$$

Розв'язавши це рівняння, ми знайдемо власні вектори, що відповідають власним значенням. Власні значення

$[\lambda = 11{,}0, \lambda = 6{,}4, \lambda = 0{,}58]$, а відповідні власні вектори будуть:

$$\begin{pmatrix} 0.63 \\ -0.48 \\ -0.62 \end{pmatrix} \begin{pmatrix} 0.24 \\ -0.64 \\ 0.73 \end{pmatrix} \begin{pmatrix} 0.74 \\ 0.61 \\ 0.28 \end{pmatrix}$$

6 крок - Сортувати та вибрати

Сортуємо власні вектори відповідно до спадання власних значень і беремо перші n власних векторів.

Власні вектори лише визначають напрямок нових вимірів у просторі.

Формуємо набір даних відповідно до цих векторів.

Метою було створення простору меншої розмірності за допомогою PCA, тому виберемо кількість власних векторів відповідно до кількості факторів, яку ми бажаємо.

Розглянемо на прикладі. Сортуємо власні значення від найбільшого до найменшого. Наприклад, якщо ви хочете зменшити розмірність до двох з трьох, то візьміть перші два власних вектора, які відповідають першим двом найбільшим власним значенням.

$$\lambda = 11, \qquad \text{власний вектор} = \begin{pmatrix} 0.63 \\ -0.48 \\ -0.62 \end{pmatrix}$$

$$\lambda = 6.4, \qquad \text{власний вектор} = \begin{pmatrix} 0.24 \\ -0.64 \\ 0.73 \end{pmatrix}$$

Кінцева мета — перетворення реального набору даних у новий простір.

Наприклад, можемо мати інформацію про всіх учнів у просторі розмірності два.

Потрібно помножити нормалізований набір даних та матрицю власних векторів, створену шляхом взяття найвищих власних значень. При множенні матриць кількість стовпців у першій матриці має дорівнювати кількості рядків у другій матриці.

$[3 \times 5]$ матриця буде помножена на (2 стовпці, 3 рядки), щоб отримати перетворений набір даних у нові розміри.

$[\text{Нормалізована матриця A}] \cdot [\text{матриця власних векторів}] =$ перетворений набір даних

$$\begin{pmatrix} 1 & -1.6 & -1.2 \\ 3 & -2.6 & -2.2 \\ 0 & 2.4 & -1.2 \\ -1 & -0.6 & 4.8 \\ -3 & 2.4 & -0.2 \end{pmatrix} \cdot \begin{pmatrix} 0.63 & 0.24 \\ -0.48 & -0.64 \\ -0.62 & 0.73 \end{pmatrix} = \begin{pmatrix} 2.13 & 0.38 \\ 4.47 & 0.76 \\ -0.40 & -2.41 \\ -3.31 & 3.66 \\ -2.90 & -2.39 \end{pmatrix}$$

Червоні квадрати вказують, який рядок множиться на який стовбець, щоб отримати перший елемент у новому просторі відповідно до власних векторів.

На завершення, ви можете побачити кінцевий набір даних для учнів у просторі розмірності два, визначених PCA, і ті самі оцінки з вихідної таблиці.

	PCA -1	PCA -2	Label
Student -1	2.13	0.38	0
Student - 2	4.47	0.76	1
Student - 3	-0.40	-2.41	1
Student - 4	-3.31	3.66	0
Student - 5	-2.90	-2.39	0

Табл. 3.2

Висновок

У роботі було досліджено метод головних компонент. У роботі наведено алгебраїчний приклад та приклад із використанням мови програмування R.

Метод головних компонент знижує розмірність датасету, тим самим є дуже корисним при дослідженні великих об’ємів даних. Використовуючи декілька головних компонент можна отримати ту ж саму інформацію, якщо використати кожен фактор окремо.

Також для реалізації цього методу можна використовувати мову програмування Python.

Перелік використаних джерел

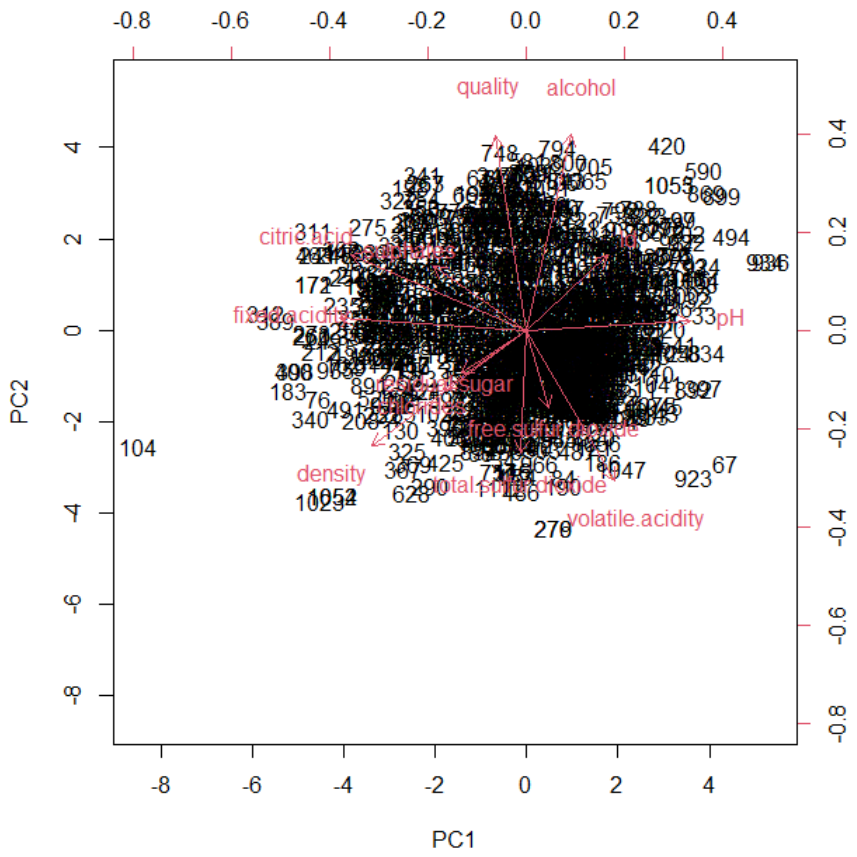
1. https://medium.com/geekculture/the-math-behind-everything-about-principle-component-analysis-pca-d6f0baff5681
2. https://uk.economy-pedia.com/11031154-vectors-and-eigenvalues
3. https://studfile.net/preview/5722599/page:3/
4. https://stud.com.ua/154764/informatika/metod_golovnih_komponent
5. http://ebooks.git-elt.hneu.edu.ua/babap/8-4-id8-4.html
6. https://pidru4niki.com/18000102/statistika/metod_golovnih_komponent_zagalne_ponyattya_metodu_yogo_zavdannya
7. https://medium.com/geekculture/dimensionality-reduction-and-principal-component-analysis-6e4d8ec6a312
8. https://www.datadriveninvestor.com/2021/04/08/the-math-behind-everything-about-principle-component-analysis-pca/

Додаток

```
library(dplyr)

dataWine <- read.table(file = "WineQT.csv", header = TRUE, sep = ",")
summary(dataWine)
str(dataWine)
#calculate principal components
results <- prcomp(dataWine, scale = TRUE)
#reverse the signs
results$rotation <- -1*results$rotation
#display principal components
results$rotation
#reverse the signs of the scores
results$x <- -1*results$x

#display the first six scores
head(results$x)
biplot(results, scale = 0)
```



	PC1	PC2	PC3	PC4	PC5
fixed.acidity	-0.47750067	0.03207913	-0.14110177	0.23229137	-0.08593363
volatile.acidity	0.22369614	-0.38324503	-0.18501857	-0.02661006	0.33018348
citric.acid	-0.44707047	0.19392824	0.11730801	0.02765048	-0.11040697
residual.sugar	-0.16608228	-0.10924891	0.24849085	0.41600742	0.69468691
chlorides	-0.21158396	-0.15183135	0.04397806	-0.64356730	0.31846579
free.sulfur.dioxide	0.05584784	-0.19774553	0.64222500	0.04234129	-0.14602576
total.sulfur.dioxide	-0.01369381	-0.31130712	0.57703504	0.04804604	-0.21690955
density	-0.39503221	-0.29272457	-0.13119569	0.19681472	0.18187621
pH	0.41520442	0.02396455	0.02673612	0.01736526	0.23215507
sulphates	-0.23669015	0.16283452	0.20000000	-0.53858969	0.20806678
alcohol	0.11395930	0.49875547	0.16678464	0.13774642	0.26046290
quality	-0.07883684	0.49553927	0.14637357	0.06245743	0.07950882
Id	0.20677085	0.19075973	0.12579199	-0.03513367	0.12079588

	PC6	PC7	PC8	PC9	PC10
fixed.acidity	0.09062337	-0.199506255	-0.21131844	0.24355340	-0.18297638
volatile.acidity	0.03839342	-0.597016402	-0.33939048	0.16057473	-0.07547432
citric.acid	0.11986865	0.007617279	0.30701690	0.26028785	-0.21521419
residual.sugar	0.04224183	0.130820920	0.21150583	-0.20927032	0.29438819
chlorides	0.17328815	-0.128777467	0.32982403	-0.23598731	-0.38216974
free.sulfur.dioxide	0.03401032	0.001434282	-0.19281577	0.03310938	-0.33561173
total.sulfur.dioxide	-0.08008886	-0.191765715	0.04949840	0.03636895	0.15244282
density	-0.09179933	0.289701653	-0.31002356	0.11776705	-0.27837380
pH	-0.36701501	0.404673397	0.05411572	0.27917934	-0.47003428
sulphates	-0.27126326	0.159745547	-0.38761318	0.32201945	0.42199198
alcohol	-0.04765820	-0.412727665	0.20189400	0.42760324	-0.07149929
quality	-0.21246254	-0.164233181	-0.41061301	-0.59649265	-0.25262504
Id	0.82002332	0.250427318	-0.30675934	0.10530908	-0.02193085