Тех.Завдання:

Дані вимірюють споживання білка в двадцяти п’яти європейських країнах для дев’яти груп продуктів(виміряні в умов. одиницях, дані датуються початком 1970-х років). Після розміщення даних в текстовому файлі їх слід зчитати у фрейм за допомогою команди read.csv(), оскільки дані в файлі розділені комами, а не пробілами.

Об’єм вибірки:25

Назви змінних:

* Country: назва країни;
* RedMeat: рівень споживання червоного м’яса;
* WhiteMeat: рівень споживання білого м’яса;
* Eggs: рівень споживання яєць;
* Milk: рівень споживання молока;
* Fish: рівень споживання риби;
* Cereals: рівень споживання злаків;
* Starch: рівень споживання крохмалю;
* Nuts: рівень споживання бобових, горіхів і олійного насіння;
* Fr&Veg: рівень споживання овочів і фруктів.

Необхідно:

* Побудувати кореляційні матриці коефіцієнтів Пірсона, Спірмена і Кендалла для всього масиву даних. Порівняти матриці і проінтерпретувати залежність між змінними.
* Використовуючи метод головних компонент, виділити головні компоненти даних. Проінтерпретувати результат. Побудувати діаграму розсіювання навантажень. Побудувати діаграму розсіювання по значеннях головних компонент (biplot). Чи присутні окремі кластери, на які розділились спостереження ?
* За допомогою однофакторного дисперсійного аналізу дослідити: чи впливає рівень змінної RedMeat на середнє значення змінної WhiteMeat.
* Зробити висновки, базуючись на ваших дослідженнях.

Дані

Country, RedMeat, WhiteMeat, Eggs, Milk, Fish, Cereals, Starch, Nuts, Fr&Veg

Albania, 10.1, 1.4, 0.5, 8.9, 0.2, 42.3, 0.6, 5.5, 1.7

Austria, 8.9, 14.0, 4.3, 19.9, 2.1, 28.0, 3.6, 1.3, 4.3

Belgium, 13.5, 9.3, 4.1, 17.5, 4.5, 26.6, 5.7, 2.1, 4.0

Bulgaria, 7.8, 6.0, 1.6, 8.3, 1.2, 56.7, 1.1, 3.7, 4.2

Czechoslovakia, 9.7, 11.4, 2.8, 12.5, 2.0, 34.3, 5.0, 1.1, 4.0

Denmark, 10.6, 10.8, 3.7, 25.0, 9.9, 21.9, 4.8, 0.7, 2.4

E\_Germany, 8.4, 11.6, 3.7, 11.1, 5.4, 24.6, 6.5, 0.8, 3.6

Finland, 9.5, 4.9, 2.7, 33.7, 5.8, 26.3, 5.1, 1.0, 1.4

France, 18.0, 9.9, 3.3, 19.5, 5.7, 28.1, 4.8, 2.4, 6.5

Greece, 10.2, 3.0, 2.8, 17.6, 5.9, 41.7, 2.2, 7.8, 6.5

Hungary, 5.3, 12.4, 2.9, 9.7, 0.3, 40.1, 4.0, 5.4, 4.2

Ireland, 13.9, 10.0, 4.7, 25.8, 2.2, 24.0, 6.2, 1.6, 2.9

Italy, 9.0, 5.1, 2.9, 13.7, 3.4, 36.8, 2.1, 4.3, 6.7

Netherlands, 9.5, 13.6, 3.6, 23.4, 2.5, 22.4, 4.2, 1.8, 3.7

Norway, 9.4, 4.7, 2.7, 23.3, 9.7, 23.0, 4.6, 1.6, 2.7

Poland, 6.9, 10.2, 2.7, 19.3, 3.0, 36.1, 5.9, 2.0, 6.6

Portugal, 6.2, 3.7, 1.1, 4.9, 14.2, 27.0, 5.9, 4.7, 7.9

Romania, 6.2, 6.3, 1.5, 11.1, 1.0, 49.6, 3.1, 5.3, 2.8

Spain, 7.1, 3.4, 3.1, 8.6, 7.0, 29.2, 5.7, 5.9, 7.2

Sweden, 9.9, 7.8, 3.5, 24.7, 7.5, 19.5, 3.7, 1.4, 2.0

Switzerland, 13.1, 10.1, 3.1, 23.8, 2.3, 25.6, 2.8, 2.4, 4.9

UK, 17.4, 5.7, 4.7, 20.6, 4.3, 24.3, 4.7, 3.4, 3.3

USSR, 9.3, 4.6, 2.1, 16.6, 3.0, 43.6, 6.4, 3.4, 2.9

W\_Germany, 11.4, 12.5, 4.1, 18.8, 3.4, 18.6, 5.2, 1.5, 3.8

Yugoslavia, 4.4, 5.0, 1.2, 9.5, 0.6, 55.9, 3.0, 5.7, 3.2

1) Прибрав стовпчик Country за допомогою d <- data[ , -1], тому що ця змінна не корелює

Обчислив матриці за допомогою функції cor



З пакету ggplot2 за допомогою функції ggcore побудував матриці



**Вигляд матриць**





Інтерпретація кореляційних матриць

* Кореляційна матриця Пірсона:
* Сильна негативна кореляція між Eggs і Cereals (-0,7), що означає, що в країнах з високим споживанням яєць низьке споживання злаків.
* Позитивна кореляція між Nuts і Cereals (0.7), що свідчить про те, що країни з високим споживанням горіхів мають тенденцію до високого споживання злаків.
* Кореляційна матриця Спірмена:
* Сильна негативна кореляція між Eggs і Cereals (-0.7), подібна до матриці Пірсона. На цей рівень також піднялась кореляція між молоком і пластівцями.
* Подібно до Пірсона найвища позитивна кореляція між Nuts і Cereals (0.7) у цій вибірці даних.
* Кореляційна матриця Кендалла:
* Загалом тенденції кореляції зберігаються, тільки дані більш розсіяні біля 0, вказуючи на слабку залежність.

2) Знову беремо змінну d (без country) , розраховуємо та будуємо графік щоб визначити основні компоненти



Графік:



Бачимо сильний вплив першої компоненти(набагато більший порівняно з іншими) на дисперсію.

Далі, будуємо діаграми розсіювання навантажень та розсіювання по значеннях головних компонент(biplot)



Діаграма розсіювання



Будуємо biplot і одразу бачимо кластер



3)За допомогою однофакторного дисперсійного аналізу досліджуємо чи впливає рівень змінної RedMeat на середнє значення змінної White Meat.

Для дисперсійного аналізу використав метод ANOVA(analysis of varience) - aov() в R



Результат:



Pr(>F) (p-значення): Це ймовірність отримати значення F-статистики, що рівне або більше за фактичне значення, якщо нульова гіпотеза (відсутність впливу) справджується. Зазвичай, якщо це значення менше за 0.05, то ми вважаємо вплив статистично значущим.

У нашому випадку, p-значення для RedMeat дорівнює 0.465, що більше за 0.05. Це означає, що немає статистично значущого впливу рівня споживання червоного м'яса на середнє значення споживання білого м'яса.

4) Висновки

1. Кореляційний аналіз

Кореляційні матриці Пірсона, Спірмена та Кендалла показують схожі тенденції у взаємозв’язках між змінними. Найбільш помітними є сильна негативна кореляція між споживанням яєць (Eggs) та злаків (Cereals), а також позитивна кореляція між споживанням горіхів (Nuts) та злаків (Cereals).

Це свідчить про те, що в країнах з високим споживанням яєць зазвичай низьке споживання злаків, і навпаки. Водночас країни з високим споживанням горіхів також мають тенденцію до високого споживання злаків.

2. Аналіз головних компонент (PCA):

Результати PCA показують, що перша головна компонента значно впливає на загальну дисперсію даних, тобто вона пояснює найбільшу частину варіації в даних.

Діаграма розсіювання навантажень та biplot показують, що деякі країни формують окремі кластери, що вказує на певні подібності в їхніх харчових уподобаннях. Це може бути використано для виявлення груп країн з схожими харчовими звичками.

3. Однофакторний дисперсійний аналіз (ANOVA):

Результати ANOVA показують, що p-значення для змінної RedMeat дорівнює 0.465, що значно більше за 0.05.

Це означає, що рівень споживання червоного м’яса (RedMeat) не має статистично значущого впливу на середнє значення споживання білого м’яса (WhiteMeat). Відповідно, ми не можемо відхилити нульову гіпотезу про відсутність впливу.

4. Загальні висновки:

- Хоча деякі змінні мають суттєві кореляційні зв’язки, рівень споживання червоного м’яса не впливає на споживання білого м’яса.

- Аналіз головних компонент дозволяє виявити ключові фактори, що впливають на харчові звички, і ідентифікувати групи країн з подібними профілями споживання.