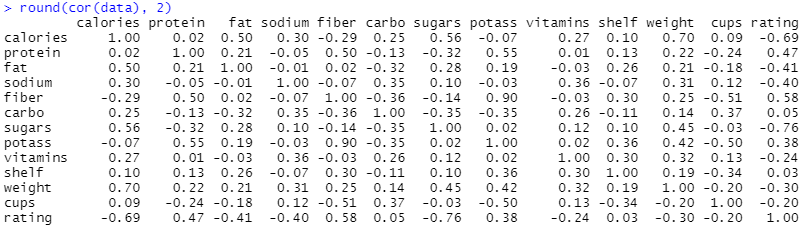
**Лабораторна робота № 8**

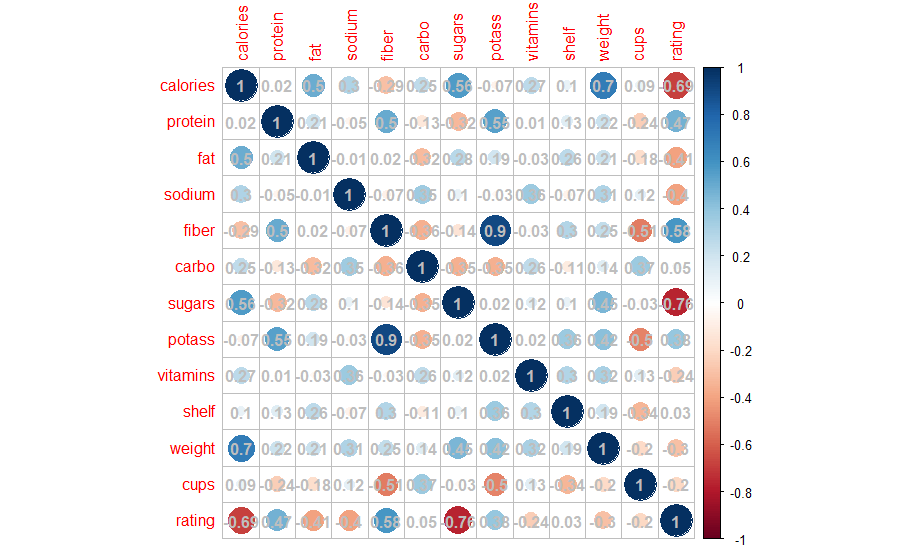
**Завдання 1: Перевірити дані на мультиколінеарність.**

1. **Представити залежність таблично (round(cor(data), 2))**

****

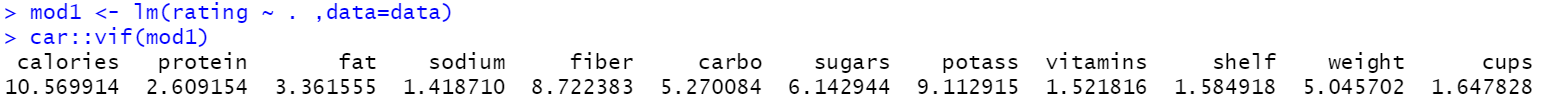
*Найбільші значення мають calories і weight(0.7) та potass і fiber(0.9). На залежну змінну rating дуже впливають змінні calories, sugar.*

1. **Представити залежність графічно (corrplot::corrplot(cor(wine), addCoef.col = "grey")).**

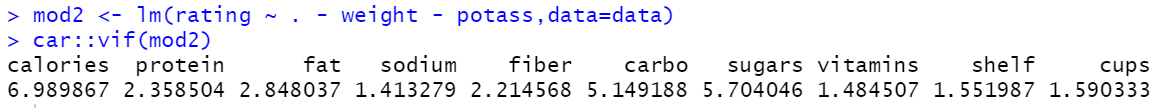
****

*З графічного представлення гарно видно, що найбільші значення мають calories і weight(0.7) та potass і fiber(0.9).*

1. **За допомогою обчислення коефіцієнта Variance Inflation Factor (VIF) перевірити змінні на мультиколінеарність. Рекомендується видалити фактор з показником VIF, який вказує на мультиколінеарність.**

****

*Великі значення коефіцієнта Variance Inflation Factor (VIF) мають такі змінні: calories, fiber, potass, weight. Оскільки fiber і potass сильно корелюють і potass має більший VIF коефіцієнт, то можемо видалити змінну potass з моделі. Між calories і weight, варто видалити weight оскільки вона має менший вплив на залежну змінну.*

****

*Після видалення змінних weight і potass, значення коефіцієнтів зменшилися, однак коефіцієнт calories все одно великий. Проте видалення його чи інших змінних (окрім незначущих) привиде до сильного пониження adjusted R^2 та F критерію.*

1. **Порівняти моделі mod\_1(y~x1+x2+x3+x4), з мультиколінеарністю, та mod\_2(y~x1+x2+x3), без мультиколінеарності, використовуючи car::compareCoefs(mod\_1, mod\_2), confint(\*) та summary(\*).**

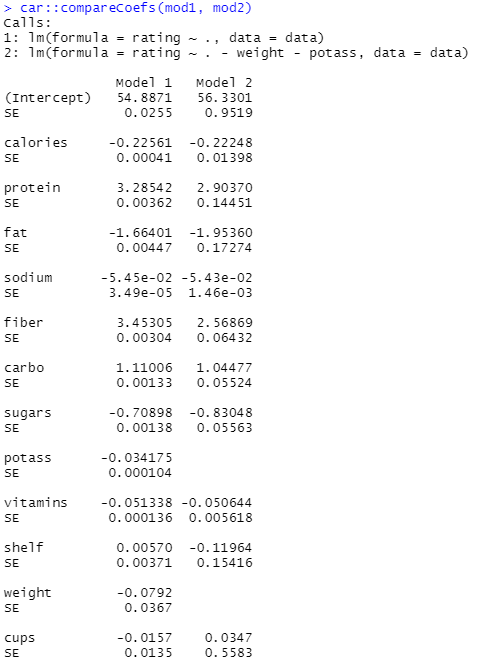
**Зображення, що містить стіл

Автоматично згенерований опис**

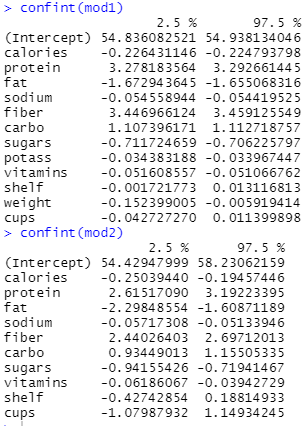
**Зображення, що містить стіл

Автоматично згенерований опис**

*З summery видно, що F статистика, adjusted R^2 і RSE краще у перший моделі з мультиколінеарністю .*

****

*Якщо порівнювати коефіцієнти, то видно, що у всіх змінних з першої моделі похибка менше.*

****

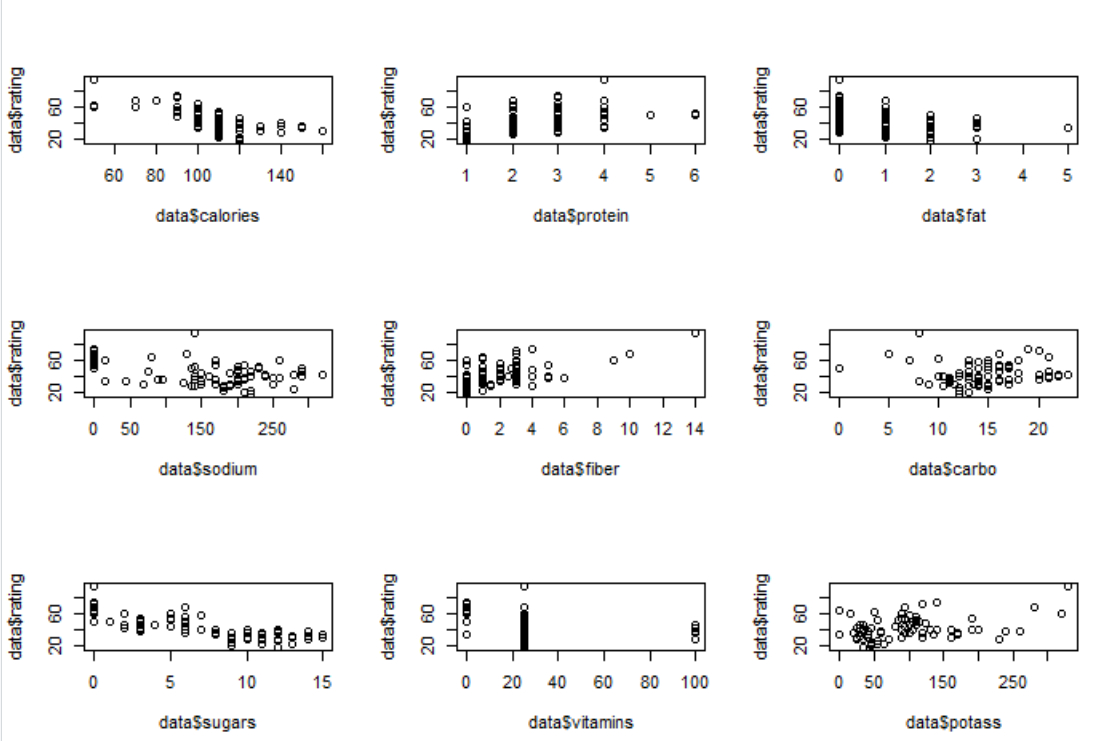
*Обидві моделі мають незначущі коефіцієнти cups і shelf.*

1. **Зробити висновки, яка модель краща mod\_1 чи mod\_2.**

*По всім параметрам краще модель перша.*

**Завдання 2: Перевірити дані на гомоскедастичність.**

1. **Графічно представити залежність між залежною та незалежними змінними (plot(data$y, data$x\_i));**

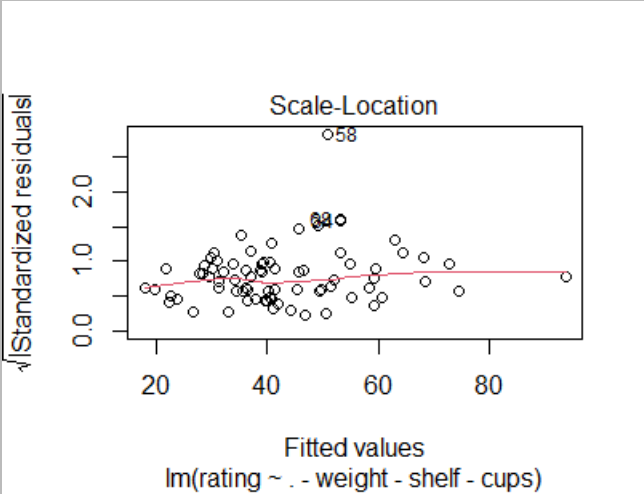
****

1. **Для перевірки на гомоскедастичність використати тест Брейша-Пагана (car::ncvTest(mod));**

**Зображення, що містить текст

Автоматично згенерований опис**

*Оскільки p-value = 0,044612 < 0,05, то це означає що присутня слабка гетероскедастичність.*

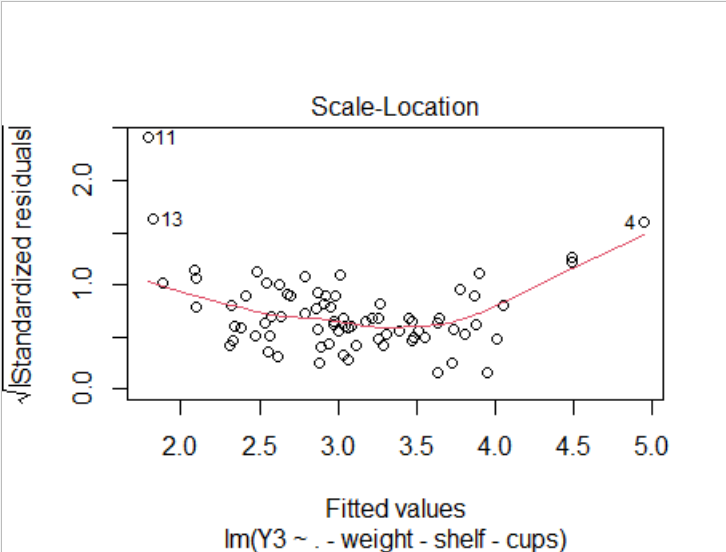
1. **Перевірити дані на гомоскедастичність за допомогою графічного методу plot(\*, 3)**

*Червона лінія розташована горизонтально, що добре, проте значення не однаково розподілені, що може вказувати на те що гетероскедастичність присутня.*

1. **Виконати перетворення для залежної змінної Y1 <- log(abs(Y)) та Y2 <- sqrt(abs(Y)). Порівняти за тестом Брейша-Пагана моделі із залежними змінними Y1 та Y2;  
   Зображення, що містить текст

   Автоматично згенерований опис**

*Бачимо що у моделі де Y1 <- log(abs(Y)) значення p-value набагато більше, тому це перетворення краще. Обидва перетворення мають p-value > 0.05, що означає, що залишки є гомоскедастичними.*

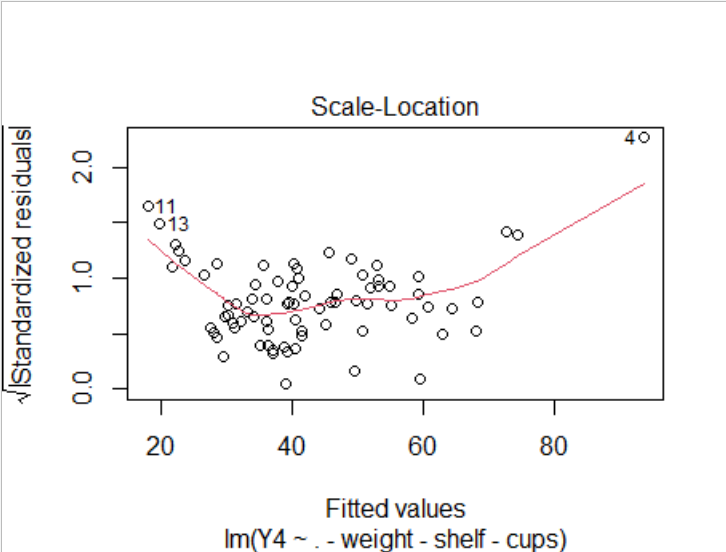
1. **Трансформація Бокса-Кокса за допомогою зсуву Y3 <- log(Y + m)). Порівняти за тестом Брейша-Пагана модель із трансформованою залежною змінною Y3;**

**Зображення, що містить текст

Автоматично згенерований опис**

*Це перетворення спрацювало гірше ніж попередні, тому що за тестом ми маємо дуже маленьке значення p-value і це означає, що присутня сильна гетероскедастичність.*

1. **Трансформація за Йо-Джонсоном Y4. Порівняти за тестом Брейша-Пагана модель із трансформованою залежною змінною Y4;**



**Зображення, що містить текст

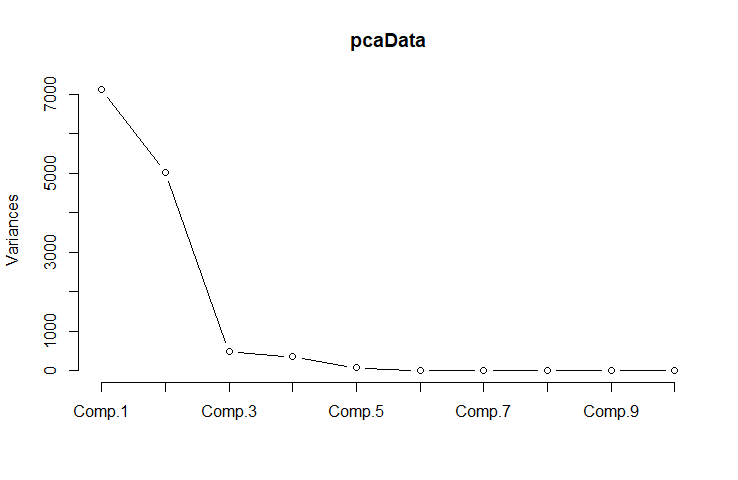
Автоматично згенерований опис**

*Це перетворення краще ніж Бокса-Кокса, проте однаково p-value<0.05 і це означає, що присутня гетероскедастичність*

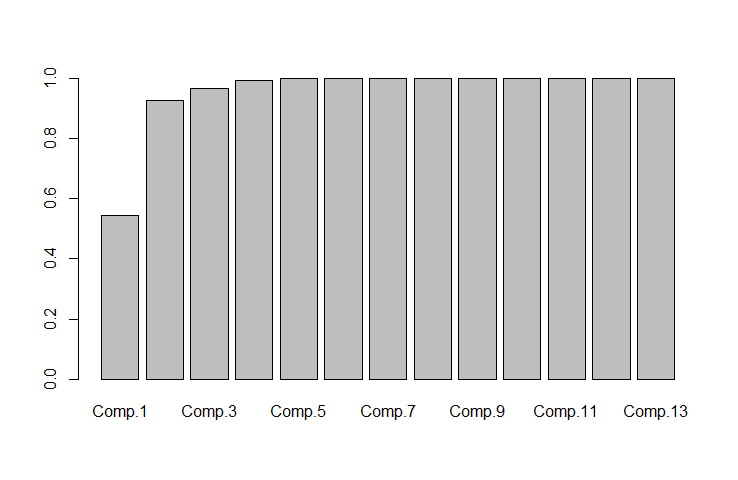
**Завдання 3: Метод головних компонент (Principal Component Analysis – PCA).**

1. **Підготовка до методу PCA (всі змінні мають тип num);  
   A picture containing text

   Description automatically generated**
2. **Метод PCA тобто princomp(data, fix\_sign = TRUE);  
   A picture containing calendar

   Description automatically generated**
3. **Діаграма дисперсій кожної компоненти plot(mod\_pca, type = "l"). Зробити висновок про кількість основних компонент, які варто брати до уваги;  
   **

*За діаграмою можна зробити висновок, що варто брати 5 компонент оскільки далі усі однакові.*

1. **Aльтернативна діаграма сукупної відсоткової дисперсії barplot(cumsum(mod\_pca$sdev^2) / sum(mod\_pca$sdev^2));**

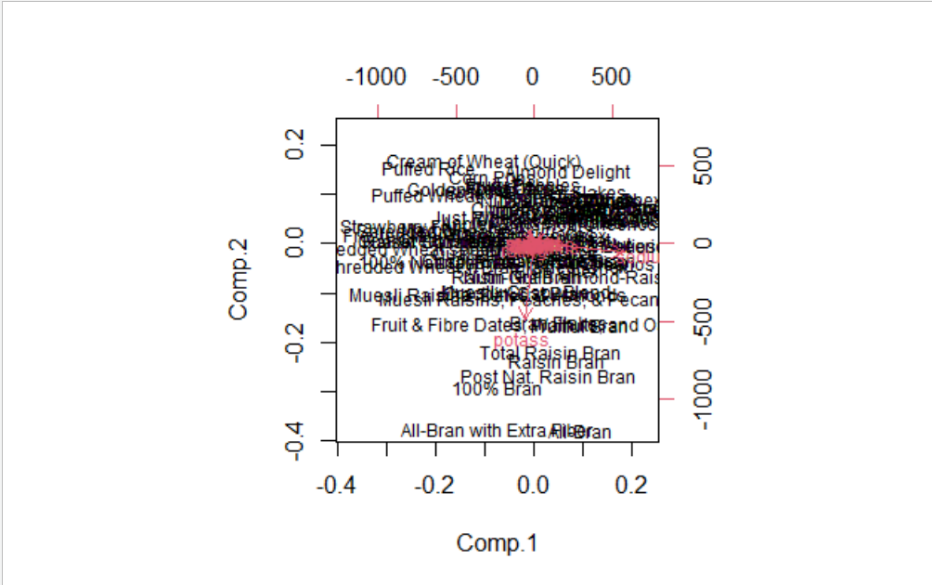
*Альтернативна діаграм підтверджує, що варто брати з 1 по 5 компоненти*

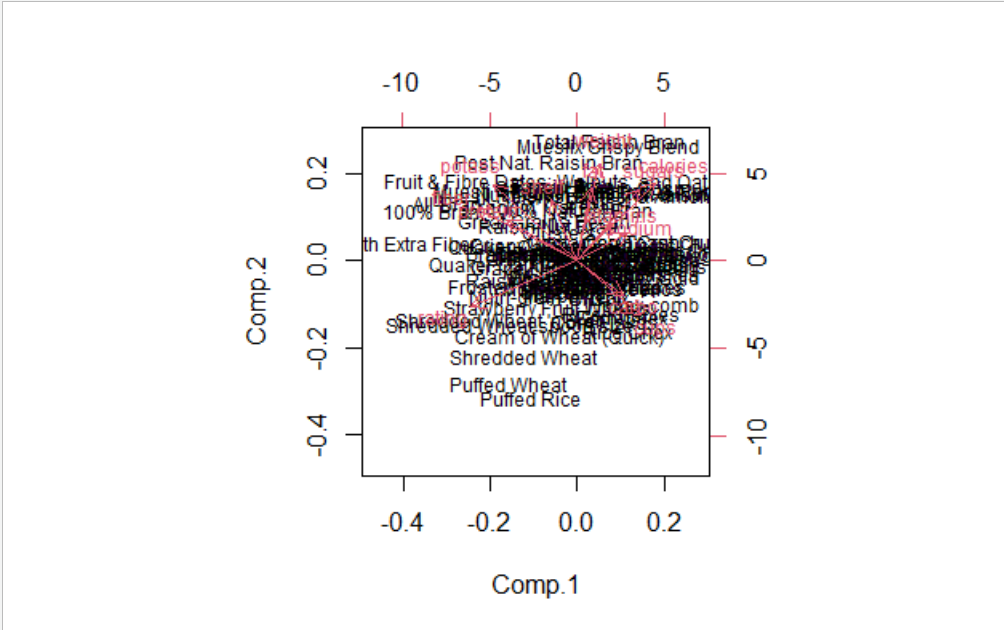
1. **Відновлення даних з усіх основних компонентів;  
   Зображення, що містить стіл

   Автоматично згенерований опис**
2. **Метод PCA для стандартизованих змінних princomp(x = laliga, cor = TRUE, fix\_sign = TRUE)  
   Scatter chart

   Description automatically generated**
3. **Графічне подання змінних через 2-ві перші основні компоненти для звичайних даних та стандартизованих biplot(\*, cex = 0.75);**

**Звичайні дані:**

 **Стандартизовані:**



*Для стандартизованих даних добре зрозуміло, як розподілені стрілочки, а для звичайних -ні.*

1. **Для Data\_X<- subset(data, select = -c(Y)) створити основні компоненти pca\_data\_X <- princomp(x = Data\_X, cor = TRUE, fix\_sign = TRUE). За 2-ма компонентами побудувати модель modPCA <- lm(Y ~ Comp.1 + Comp.2, data).**

***Зображення, що містить текст

Автоматично згенерований опис***

***Зображення, що містить текст, чек, знімок екрана

Автоматично згенерований опис***

*Коефіцієнти при компонентах є значущі, модель має нормальне значення R-squared і F-statistic. Для більшого значення R-squared і F-statistic варто додати більше компонент( по 5).*