

Лабораторна робота № 8

Завдання 1: Перевірити дані на мультиколінеарність.

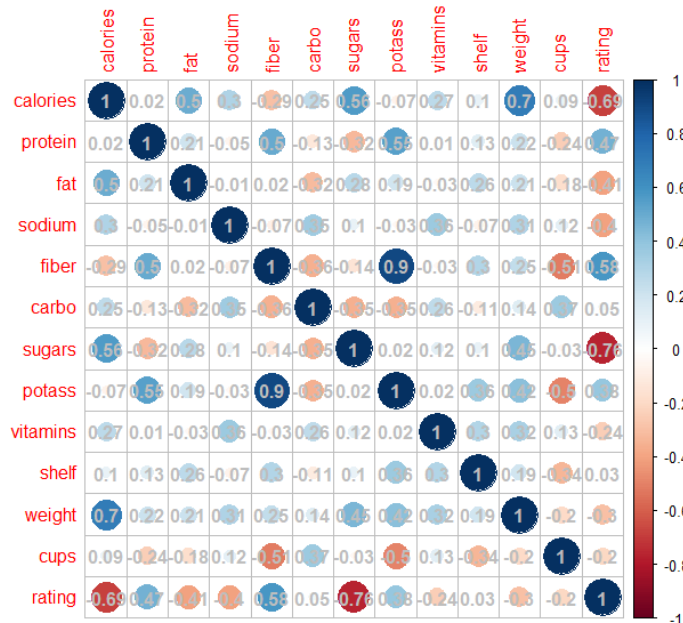
(A) Представити залежність таблично (round(cor(data), 2))

```
> round(cor(data), 2)
```

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
calories	1.00	0.02	0.50	0.30	-0.29	0.25	0.56	-0.07	0.27	0.10	0.70	0.09	-0.69
protein	0.02	1.00	0.21	-0.05	0.50	-0.13	-0.32	0.55	0.01	0.13	0.22	-0.24	0.47
fat	0.50	0.21	1.00	-0.01	0.02	-0.32	0.28	0.19	-0.03	0.26	0.21	-0.18	-0.41
sodium	0.30	-0.05	-0.01	1.00	-0.07	0.35	0.10	-0.03	0.36	-0.07	0.31	0.12	-0.40
fiber	-0.29	0.50	0.02	-0.07	1.00	-0.36	-0.14	0.90	-0.03	0.30	0.25	-0.51	0.58
carbo	0.25	-0.13	-0.32	0.35	-0.36	1.00	-0.35	-0.35	0.26	-0.11	0.14	0.37	0.05
sugars	0.56	-0.32	0.28	0.10	-0.14	-0.35	1.00	0.02	0.12	0.10	0.45	-0.03	-0.76
potass	-0.07	0.55	0.19	-0.03	0.90	-0.35	0.02	1.00	0.02	0.36	0.42	-0.50	0.38
vitamins	0.27	0.01	-0.03	0.36	-0.03	0.26	0.12	0.02	1.00	0.30	0.32	0.13	-0.24
shelf	0.10	0.13	0.26	-0.07	0.30	-0.11	0.10	0.36	0.30	1.00	0.19	-0.34	0.03
weight	0.70	0.22	0.21	0.31	0.25	0.14	0.45	0.42	0.32	0.19	1.00	-0.20	-0.30
cups	0.09	-0.24	-0.18	0.12	-0.51	0.37	-0.03	-0.50	0.13	-0.34	-0.20	1.00	-0.20
rating	-0.69	0.47	-0.41	-0.40	0.58	0.05	-0.76	0.38	-0.24	0.03	-0.30	-0.20	1.00

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

(B) 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).

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

```
> mod1 <- lm(rating ~ . , data=data)
```

```
> car::vif(mod1)
```

calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups
10.569914	2.609154	3.361555	1.418710	8.722383	5.270084	6.142944	9.112915	1.521816	1.584918	5.045702	1.647828

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.

```
> mod2 <- lm(rating ~ . - weight - potass, data=data)
> car::vif(mod2)
calories protein fat sodium fiber carbo sugars vitamins shelf cups
6.989867 2.358504 2.848037 1.413279 2.214568 5.149188 5.704046 1.484507 1.551987 1.590333
```

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

(D) Порівняти моделі $\text{mod}_1(y \sim x_1 + x_2 + x_3 + x_4)$, з мультиколінеарністю, та $\text{mod}_2(y \sim x_1 + x_2 + x_3)$, без мультиколінеарності, використовуючи `car::compareCoefs(mod_1, mod_2)`, `confint(*)` та `summary(*)`.

```
lm(formula = rating ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-0.076552 -0.011151 -0.000473  0.010169  0.046746

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.489e+01  2.554e-02  2148.907 <2e-16 ***
calories     -2.256e-01  4.098e-04   -550.540 <2e-16 ***
protein       3.285e+00  3.624e-03    906.678 <2e-16 ***
fat          -1.664e+00  4.474e-03   -371.935 <2e-16 ***
sodium       -5.449e-02  3.489e-05  -1561.551 <2e-16 ***
fiber         3.453e+00  3.043e-03   1134.635 <2e-16 ***
carbo         1.110e+00  1.332e-03    833.277 <2e-16 ***
sugars       -7.090e-01  1.376e-03   -515.140 <2e-16 ***
potass       -3.418e-02  1.040e-04   -328.440 <2e-16 ***
vitamins     -5.134e-02  1.356e-04   -378.589 <2e-16 ***
shelf         5.698e-03  3.714e-03     1.534  0.1299
weight       -7.916e-02  3.666e-02    -2.159  0.0346 *
cups         -1.566e-02  1.355e-02    -1.156  0.2519
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02141 on 64 degrees of freedom
Multiple R-squared:  1, Adjusted R-squared:  1
F-statistic: 2.726e+06 on 12 and 64 DF, p-value: < 2.2e-16

> summary(mod2)

Call:
lm(formula = rating ~ . - weight - potass, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-1.96924 -0.57367 -0.07974  0.48874  2.52648

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.330051  0.951921  59.175 < 2e-16 ***
calories     -0.222484  0.013979  -15.916 < 2e-16 ***
protein       2.903702  0.144514  20.093 < 2e-16 ***
fat          -1.953599  0.172740  -11.309 < 2e-16 ***
sodium       -0.054256  0.001461  -37.139 < 2e-16 ***
fiber         2.568692  0.064325  39.933 < 2e-16 ***
carbo         1.044772  0.055236  18.915 < 2e-16 ***
sugars       -0.830484  0.055630  -14.929 < 2e-16 ***
vitamins     -0.050644  0.005618   -9.015 4.1e-13 ***
shelf        -0.119640  0.154159   -0.776  0.440
cups          0.034732  0.558264   0.062  0.951
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8981 on 66 degrees of freedom
Multiple R-squared:  0.9965, Adjusted R-squared:  0.9959
F-statistic: 1853 on 10 and 66 DF, p-value: < 2.2e-16
```

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

```
> car::compareCoefs(mod1, mod2)
Calls:
1: lm(formula = rating ~ ., data = data)
2: lm(formula = rating ~ . - weight - potass, data = data)
```

	Model 1	Model 2
(Intercept)	54.8871	56.3301
SE	0.0255	0.9519
calories	-0.22561	-0.22248
SE	0.00041	0.01398
protein	3.28542	2.90370
SE	0.00362	0.14451
fat	-1.66401	-1.95360
SE	0.00447	0.17274
sodium	-5.45e-02	-5.43e-02
SE	3.49e-05	1.46e-03
fiber	3.45305	2.56869
SE	0.00304	0.06432
carbo	1.11006	1.04477
SE	0.00133	0.05524
sugars	-0.70898	-0.83048
SE	0.00138	0.05563
potass	-0.034175	
SE	0.000104	
vitamins	-0.051338	-0.050644
SE	0.000136	0.005618
shelf	0.00570	-0.11964
SE	0.00371	0.15416
weight	-0.0792	
SE	0.0367	
cups	-0.0157	0.0347
SE	0.0135	0.5583

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

```
> confint(mod1)
                2.5 %      97.5 %
(Intercept) 54.836082521 54.938134046
calories    -0.226431146 -0.224793798
protein      3.278183564  3.292661445
fat          -1.672943645 -1.655068316
sodium       -0.054558944 -0.054419525
fiber         3.446966124  3.459125549
carbo         1.107396171  1.112718757
sugars        -0.711724659 -0.706225797
potass        -0.034383188 -0.033967447
vitamins      -0.051608557 -0.051066762
shelf         -0.001721773  0.013116813
weight        -0.152399005 -0.005919414
cups          -0.042727270  0.011399898
```

```
> confint(mod2)
                2.5 %      97.5 %
(Intercept) 54.42947999 58.23062159
calories    -0.25039440 -0.19457446
protein      2.61517090  3.19223395
fat          -2.29848554 -1.60871189
sodium       -0.05717308 -0.05133946
fiber         2.44026403  2.69712013
carbo         0.93449013  1.15505335
sugars        -0.94155426 -0.71941467
vitamins      -0.06186067 -0.03942729
shelf         -0.42742854  0.18814933
cups          -1.07987932  1.14934245
```

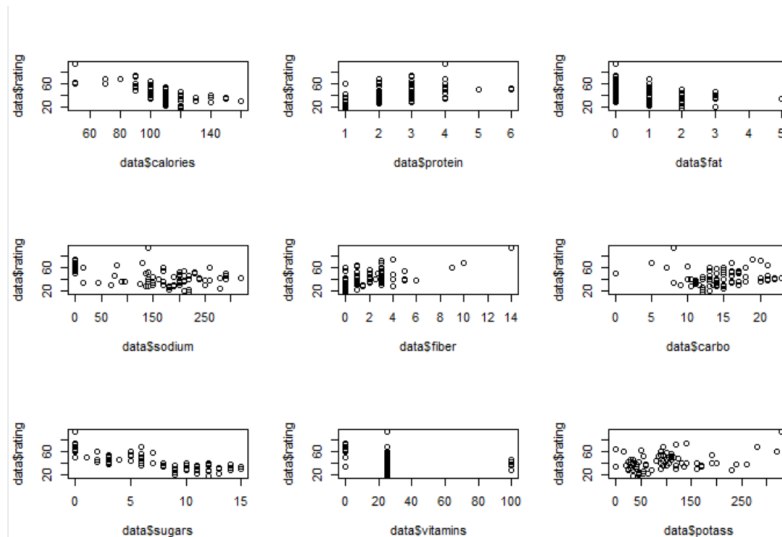
Обидві моделі мають незначущі коефіцієнти cups і shelf.

(Е) Зробити висновки, яка модель краща mod_1 чи mod_2.

По всім параметрам краще модель перша.

Завдання 2: Перевірити дані на гомоскедастичність.

(А) Графічно представити залежність між залежною та незалежними змінними (plot(data\$y, data\$x_i));



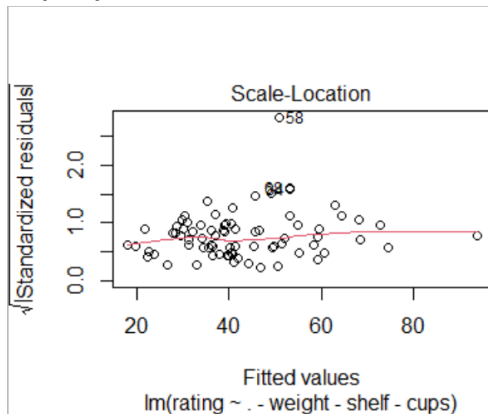
(В) Для перевірки на гомоскедастичність використати тест Брейша-Пагана (`car::ncvTest(mod)`);

```
> mod <- lm(rating ~ . - weight - shelf - cups, data=data)
> car::ncvTest(mod)
```

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.033236, Df = 1, p = 0.044612

Оскільки $p\text{-value} = 0,044612 < 0,05$, то це означає що присутня слабка гетероскедастичність.

(С) 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.

(D) Виконати перетворення для залежної змінної $Y1 <- \log(\text{abs}(Y))$ та $Y2 <- \sqrt{\text{abs}(Y)}$.

Порівняти за тестом Брейша-Пагана моделі із залежними змінними $Y1$ та $Y2$;

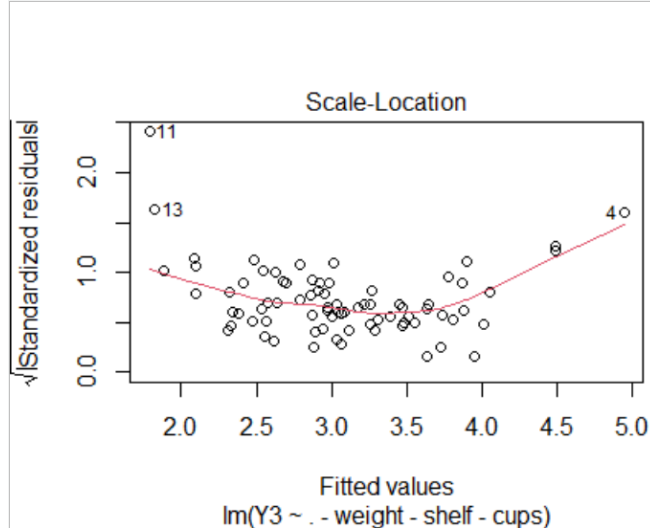
```
> Y1 <- log(abs(data$rating))
> Y2 <- sqrt(abs(data$rating))
> modY1 <- lm(Y1 ~ . - weight - shelf - cups, data=data)
> modY2 <- lm(Y2 ~ . - weight - shelf - cups, data=data)
> car::ncvTest(modY1)
```

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1001926, Df = 1, p = 0.7516

```
> car::ncvTest(modY2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.548512, Df = 1, p = 0.059599
```

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

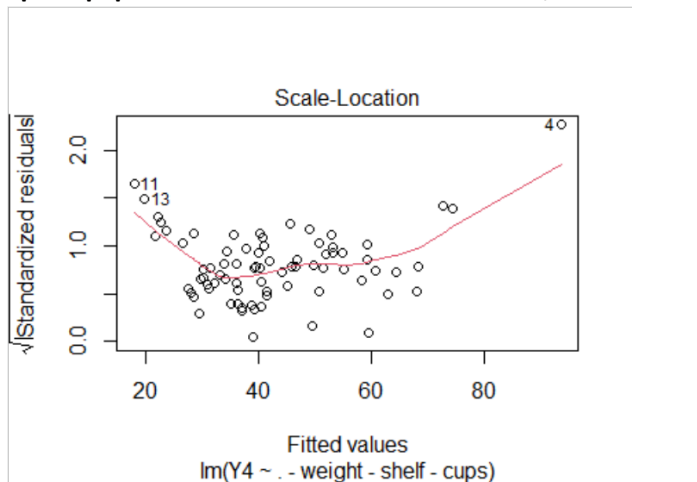
- (E) Трансформація Бокса-Кокса за допомогою зсуву $Y3 \leftarrow \log(Y + m)$. Порівняти за тестом Брейша-Пагана модель із трансформованою залежною змінною $Y3$;



```
> # Зміщений і перетворений D
> delta <- 1 # Це налаштовується
> m <- -min(data$rating) + delta
> Y3 <- log(data$rating + m)
> modY3 <- lm(Y3 ~ . - weight - shelf - cups, data=data)
> plot(modY3, 3)
> car::ncvTest(modY3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 35.78558, Df = 1, p = 2.2027e-09
```

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

- (F) Трансформація за Йо-Джонсоном $Y4$. Порівняти за тестом Брейша-Пагана модель із трансформованою залежною змінною $Y4$;



```
> # Оптимальна лямбда для Йо-Джонсона
> YJ <- car::powerTransform(mod, family = "yjPower")
> (lambdaYJ <- YJ$lambda)
YJ
0.9992163
> # Трансформація Йо-Джонсона
> Y4 <- car::yjPower(U = data$rating, lambda = lambdaYJ)
> modY4 <- lm(Y4 ~ . - weight - shelf - cups, data=data)
> plot(modY4, 3)
> car::ncvTest(modY4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 12.87157, Df = 1, p = 0.00033361
```

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).

(A) Підготовка до методу PCA (всі змінні мають тип num);

```
#### (A) ####
```

```
rownames(secondData) <- secondData$name  
secondData$name <- NULL  
secondData$mfr <- NULL  
secondData$type <- NULL
```

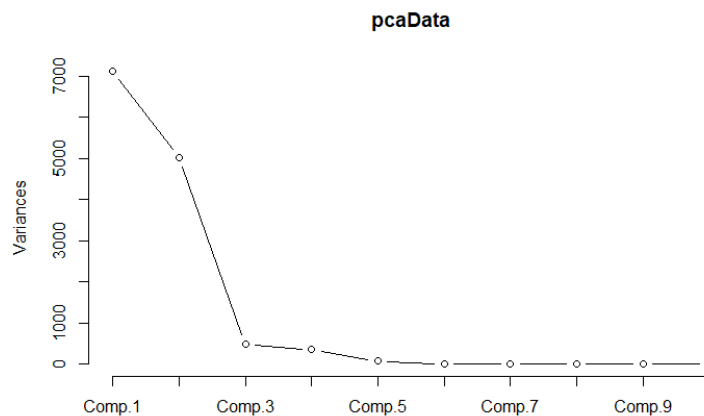
(B) Метод PCA тобто princomp(data, fix_sign = TRUE);

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Standard deviation	84.2759519	70.8716334	22.23210885	18.74352131	8.571598000	2.2697014511	2.0591916331	8.003354e-01	6.836657e-01
Proportion of Variance	0.5440361	0.3847381	0.03786009	0.02691056	0.005627869	0.0003946007	0.0003247984	4.906416e-05	3.580206e-05
Cumulative Proportion	0.5440361	0.9287742	0.96663430	0.99354486	0.999172729	0.9995673293	0.9998921277	9.999412e-01	9.999770e-01

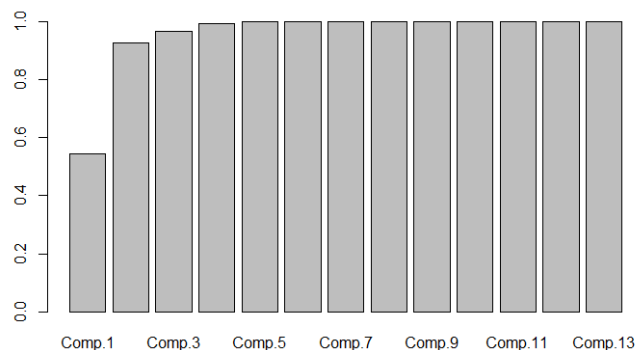
	Comp.10	Comp.11	Comp.12	Comp.13
Standard deviation	5.122623e-01	1.830804e-01	6.634331e-02	3.670147e-03
Proportion of Variance	2.010042e-05	2.567462e-06	3.371434e-07	1.031780e-09
Cumulative Proportion	9.999971e-01	9.999997e-01	1.000000e+00	1.000000e+00

(C) 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.

(D) Альтернативна діаграма сукупної відсоткової дисперсії barplot(cumsum(mod_pca\$sdev^2) / sum(mod_pca\$sdev^2));



Альтернативна діаграма підтверджує, що варто брати з 1 по 5 компоненту

(E) Відновлення даних з усіх основних компонентів;

```
> head(
+ sweep(pcaData$scores %>% t(pcaData$loadings), 2, pcaData$center, "+")
+ )
```

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf
100% Bran	70	4	1.000000e+00	130	10.0	5.0	6.00000e+00	2.80000e+02	2.50000e+01	3
100% Natural Bran	120	3	5.000000e+00	15	2.0	8.0	8.00000e+00	1.35000e+02	-1.065814e-14	3
All-Bran	70	4	1.000000e+00	260	9.0	7.0	5.00000e+00	3.20000e+02	2.50000e+01	3
All-Bran with Extra Fiber	50	4	-9.325873e-15	140	14.0	8.0	1.24345e-14	3.30000e+02	2.50000e+01	3
Almond Delight	110	2	2.000000e+00	200	1.0	14.0	8.00000e+00	1.421085e-14	2.50000e+01	3
Apple cinnamon cheerios	110	2	2.000000e+00	180	1.5	10.5	1.00000e+01	7.00000e+01	2.50000e+01	1

	weight	cups	rating
100% Bran	1	0.33	68.40297
100% Natural Bran	1	1.00	33.98368
All-Bran	1	0.33	59.42550
All-Bran with Extra Fiber	1	0.50	93.70491
Almond Delight	1	0.75	34.38484
Apple cinnamon cheerios	1	0.75	29.50954

(F) Метод PCA для стандартизованих змінних princomp(x = laliga, cor = TRUE, fix_sign = TRUE)

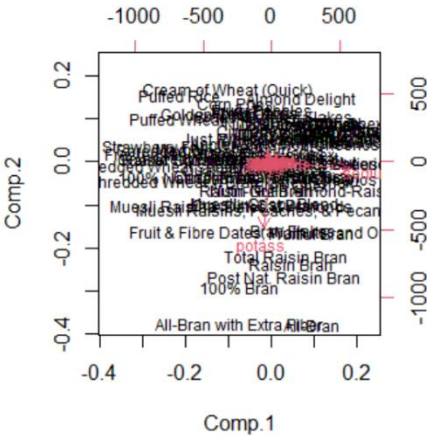
Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
standard deviation	1.899373	1.7729437	1.3516946	1.03047078	0.98583329	0.8477371	0.81993803	0.66920840	0.55108684	0.351057758
Proportion of Variance	0.277509	0.2417946	0.1405445	0.08168231	0.07475902	0.0552814	0.05171526	0.03444922	0.02336129	0.009480119
Cumulative Proportion	0.277509	0.5193036	0.6598481	0.74153036	0.81628938	0.8715708	0.92328604	0.95773527	0.98109655	0.990576670

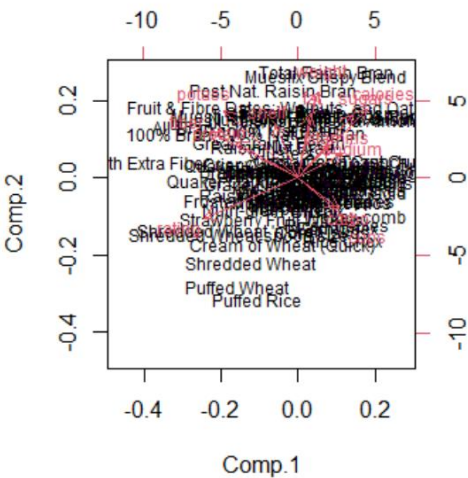
	Comp.11	Comp.12	Comp.13
standard deviation	0.269925285	0.222806101	1.035318e-03
Proportion of Variance	0.005604589	0.003818658	8.245251e-08
Cumulative Proportion	0.996181259	0.999999918	1.000000e+00

(G) Графічне подання змінних через 2-ві перші основні компоненти для звичайних даних та стандартизованих biplot(*, сех = 0.75);

Звичайні дані:



Стандартизовані:



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