Correspondence Analysis

Hyeri Goh
100349818

What is Correspondence Analysis?

Correspondence analysis is used to find out the relationship between and within two categorical variables, shown in a contingency table. Thus, we can detect similarity between columns and rows on the contingency table. We can define inertia through factors(dimensions) since the first factor represents the most of variation within the data. Correspondence analysis shrink down the dimension by clustering highly associated elements.

Contingency Table

	CEREALS	MUESLI	PORRIDGE	BACON_AND_EGGS	TOAST_AND_EGGS	FRESH_FRUIT	STEWED_FRUIT	YOGURT
Healthy	14	38	25	18	8	31	28	34
Nutritious	14	28	25	25	7	32	26	31
Good_in_Summer	42	22	11	13	7	37	16	35
Good_in_Winter	10	10	32	26	6	11	19	8
Expensive	6	33	5	27	3	9	18	10
Quick_and_Easy	54	33	8	2	15	26	8	20
Tasty	24	21	16	34	11	33	26	26
Economical	24	3	20	3	16	7	3	7
For_a_treat	5	3	3	31	4	4	16	17
For_weekdays	47	24	15	9	13	11	6	10
For_weekends	12	5	8	56	16	10	23	18
Tasteless	8	6	2	2	0	0	2	1
Takes_too_long_to_prepare	0	0	9	35	1	0	10	0
Familys_favourite	14	4	10	31	5	7	2	5

We have two categorical variables which are the type of breakfast and the perception about the type of breakfast. There are 8 kinds for the type of breakfast and 14 rows. After chi-squared test, it was concluded that H0 should be rejected since p-value is approaching to zero. Thus, there is significant dependency between two variables.

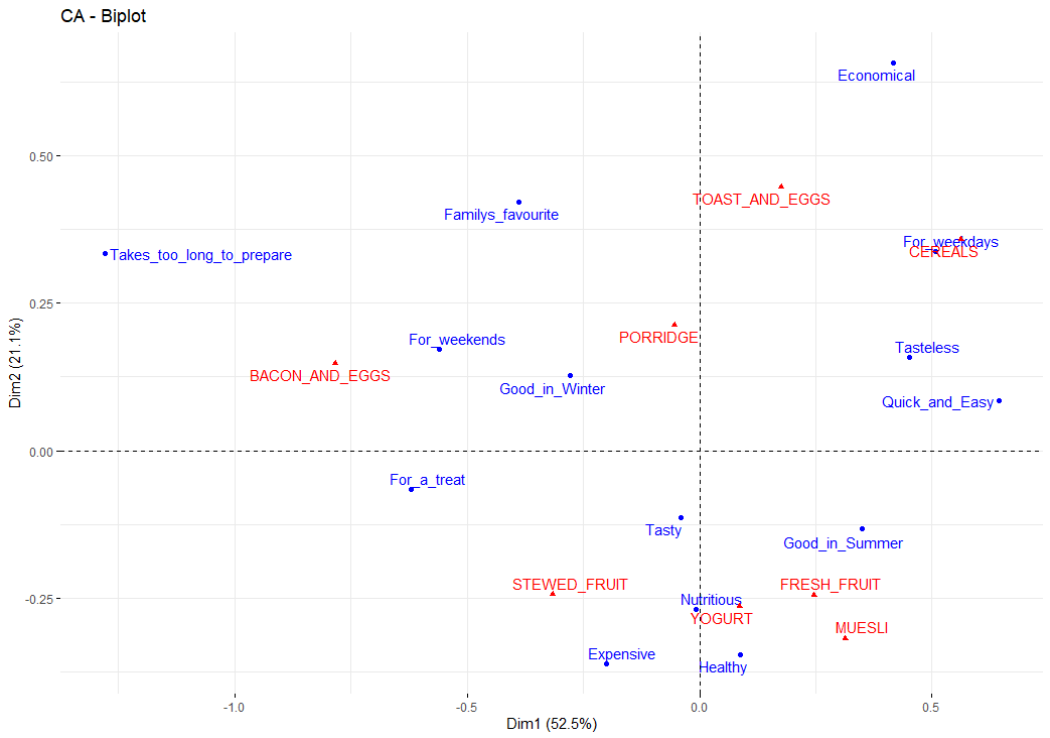
```
H0: The two categorical variables are independent.
Ha: The two categorical variables are dependent.


> chisq

        Pearson's Chi-squared test

data:  meal
X-squared = 647.31, df = 91, p-value < 2.2e-16
```

Symmetric Plot



This biplot including both rows and columns shows global pattern of two variables. As we can see from the symmetric plot, dimensions 1 and 2 explain approximately 52.5% and 21.1% of the total inertia respectively. So, with two axes, total 73.6% of variation within dataset can be explained.

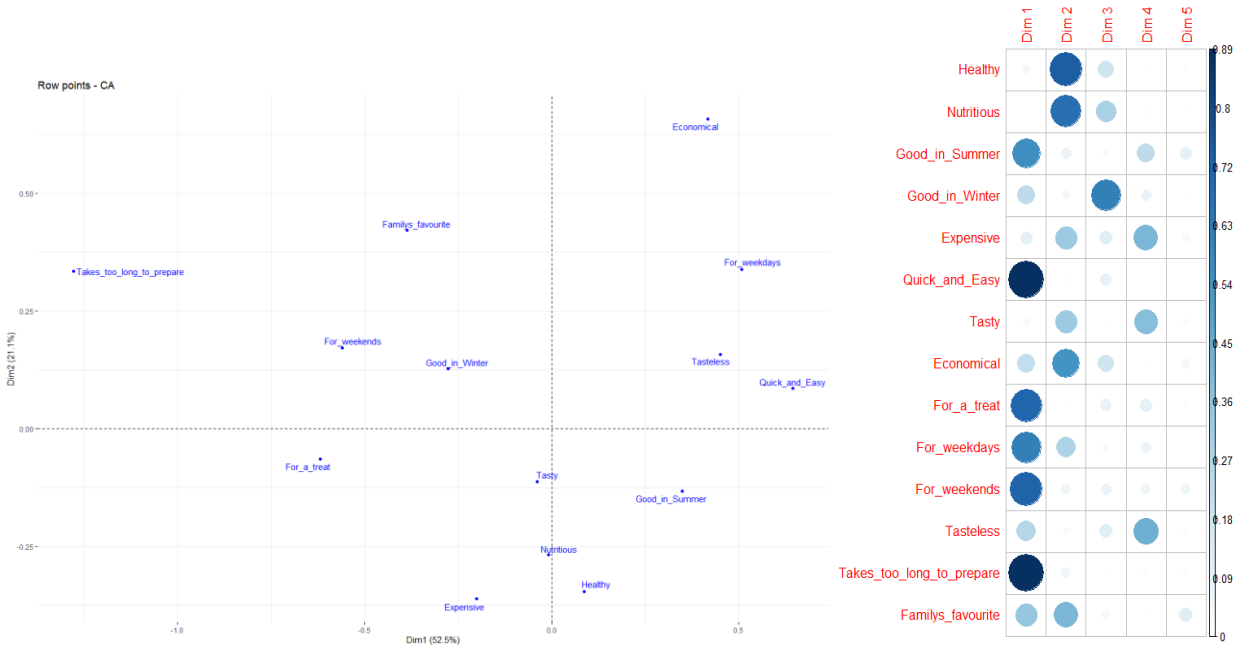
The general pattern we can read from the plot is that STEWED_FRUIT, YOGURT, FRESH_FRUIT, and MUESLI are located closed to one another. These types of breakfast are also associated with the perception of Tasty, Healthy, Nutritious, Expensive, and Good_in_Summer, which makes sense.

Also, CEREALS is very close to feature named For_weekdays and more or less associated with perception of Tasteless and Quick_and_Easy, which represent the characteristics of cereals. However, BACON_AND_EGGS and PORRIDGE are more associated with For_weekends and Good_in_Winter, which means these two types are opposite to CEREALS or FRUIT, YOGURT, MUESLI.

Asymmetric Plotw(rows)

The plot below is showing only rows points which mean the perception about the breakfast kinds. Row$coord indicates where each perception is located on each dimension. Also, row$cos2 shows the quality of representation of rows. Thus, we can interpret how much each

row represents the dimensions. For the first dimension, Takes_too_long_to_prepare, Quick_and_Easy, For_weekends, For_a_treat, and For_weekdays have good representation. For the second dimension, Healthy and Nutritious have good quality of representation.



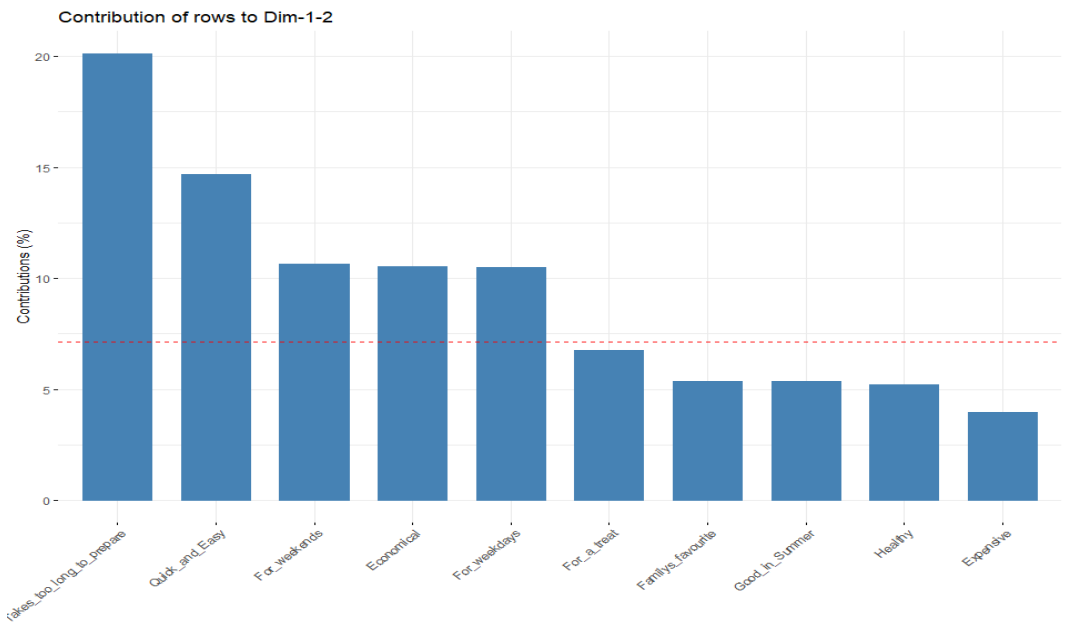

```
> row$cos2
                              Dim 1       Dim 2        Dim 3        Dim 4        Dim 5
Healthy                0.0459955559 0.733508260 1.825766e-01 0.0077393321 0.0156529061
Nutritious             0.0006932607 0.676474546 2.937468e-01 0.0141445255 0.0067662083
Good_in_Summer         0.5520848000 0.079977099 3.279771e-02 0.2320317556 0.0965732178
Good_in_Winter         0.2267089029 0.047609611 6.165307e-01 0.0781636084 0.0127459549
Expensive              0.0995136175 0.322496125 1.187374e-01 0.4103129755 0.0369501716
Quick_and_Easy         0.8887168526 0.015410243 9.052749e-02 0.0007894068 0.0002656824
Tasty                  0.0405826086 0.328360287 7.207477e-05 0.3776743617 0.0202302329
Economical             0.2161413710 0.536509623 1.818482e-01 0.0123574491 0.0507789051
For_a_treat            0.7074759278 0.007769511 8.647522e-02 0.0965955494 0.0173013878
For_weekdays           0.6108677667 0.271740622 3.502200e-02 0.0759272694 0.0001487518
For_weekends           0.7244272057 0.067721547 7.258646e-02 0.0669758451 0.0645784950
Tasteless              0.2577145243 0.031548422 1.183803e-01 0.4375380437 0.0203630999
Takes_too_long_to_prepare 0.8943759359 0.061110899 1.176126e-03 0.0227004066 0.0188809224
Familys_favourite      0.3426319857 0.405321787 4.381260e-02 0.0006825504 0.1207684885
```

Rows$contrib literally means the amount of contribution each row make for certain dimension.

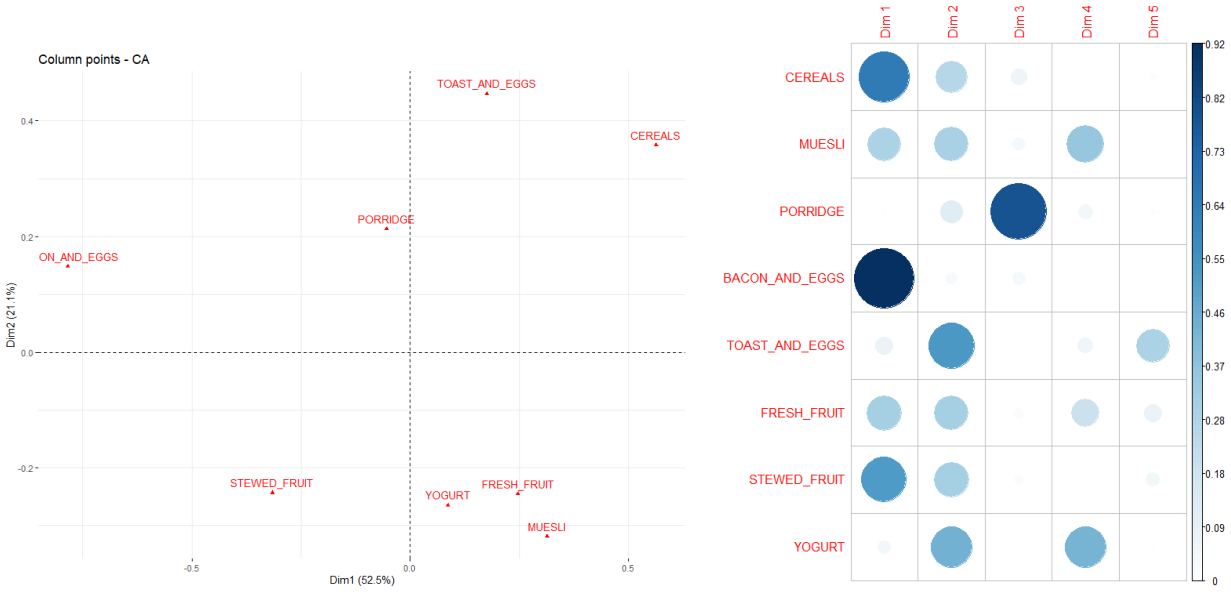
```
                              Dim 1       Dim 2
Healthy                0.432728232 17.1427781
Nutritious             0.004098089  9.9337456
Good_in_Summer         6.550654749  2.3573355
Good_in_Winter         2.762627531  1.4412032
Expensive              1.315260585 10.5884071
Quick_and_Easy        20.236188821  0.8716691
Tasty                  0.088274556  1.7742836
Economical             4.245329506 26.1774811
For_a_treat            9.371823058  0.2556719
For_weekdays          10.201065821 11.2727465
For_weekends          13.659368051  3.1720466
Tasteless              1.256632548  0.3821412
Takes_too_long_to_prepare 26.423537920  4.4850405
Familys_favourite      3.452410533 10.1454500
```

Takes_too_long_to_prepare and Quick_and_Easy contribute the most to dimension 1 while Economical contributes the most to dimension 2.

It can be shown in bar plot to find out which rows will contribute more to both of dimension.

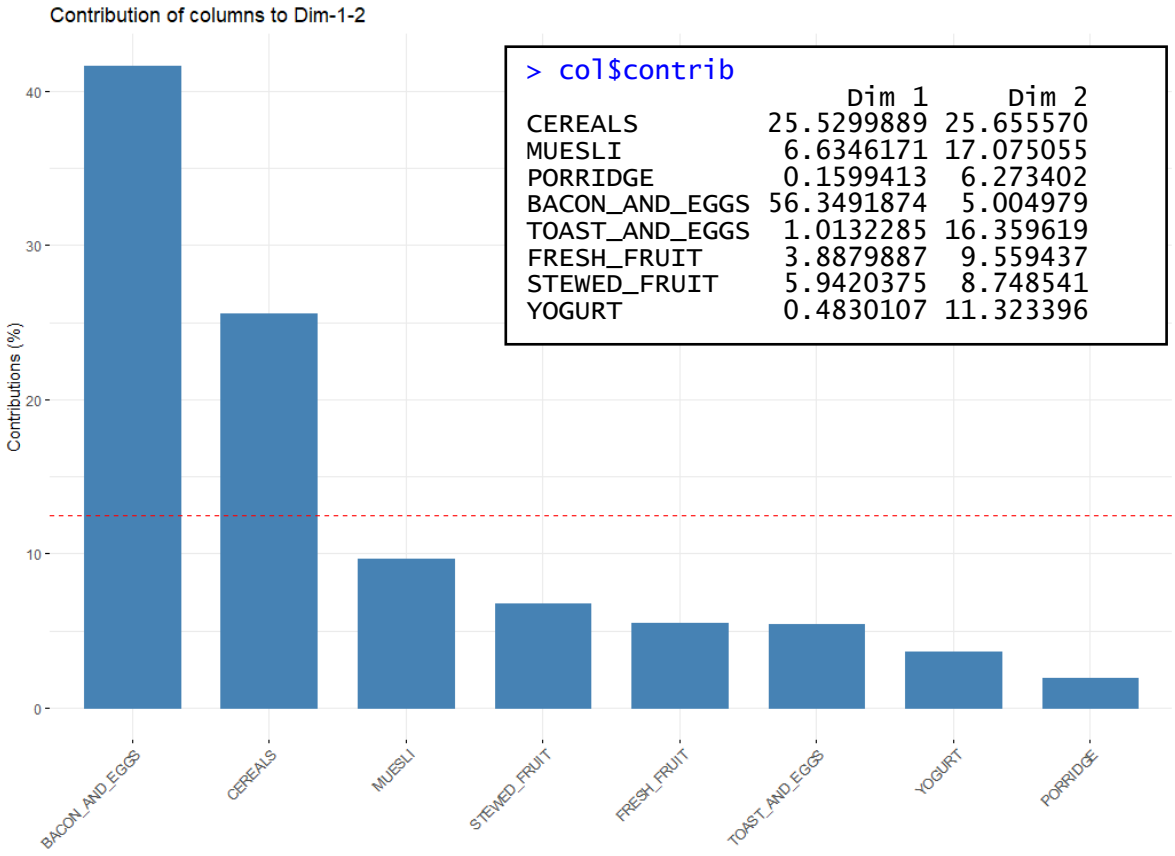


Asymmetric Plot(columns)




```
> col$cos2
                    Dim 1      Dim 2       Dim 3        Dim 4       Dim 5
CEREALS         0.646645481 0.26158959 0.070606632 0.0001489337 0.011899907
MUESLI          0.288961158 0.29937042 0.038631882 0.3572319776 0.008361719
PORRIDGE        0.008112616 0.12809334 0.796147460 0.0535261421 0.010957760
BACON_AND_EGGS  0.916602843 0.03277321 0.043722721 0.0009250488 0.004246340
TOAST_AND_EGGS  0.081455730 0.52943248 0.005955790 0.0593330843 0.288019433
FRESH_FRUIT     0.308381373 0.30522398 0.026095647 0.1970657821 0.078564498
STEWED_FRUIT    0.520264972 0.30835288 0.020517684 0.0002324215 0.048915787
YOGURT          0.046569118 0.43948201 0.001460297 0.4283951969 0.007405051
```

According to the asymmetric plot on column space, YOGURT, MUESLI, FRESH_FRUIT, and STEWED_FRUIT are gathered because they are somehow sharing similarities. Apart from these four breakfast kinds, leftover are quite distant from one another, which means they are not much similar. Also, cos2 values indicate BACON_AND_EGGS has good representation of dimension1 while TOAST_AND_EGGS has the best representation on dimension2 among eight kinds of breakfast.

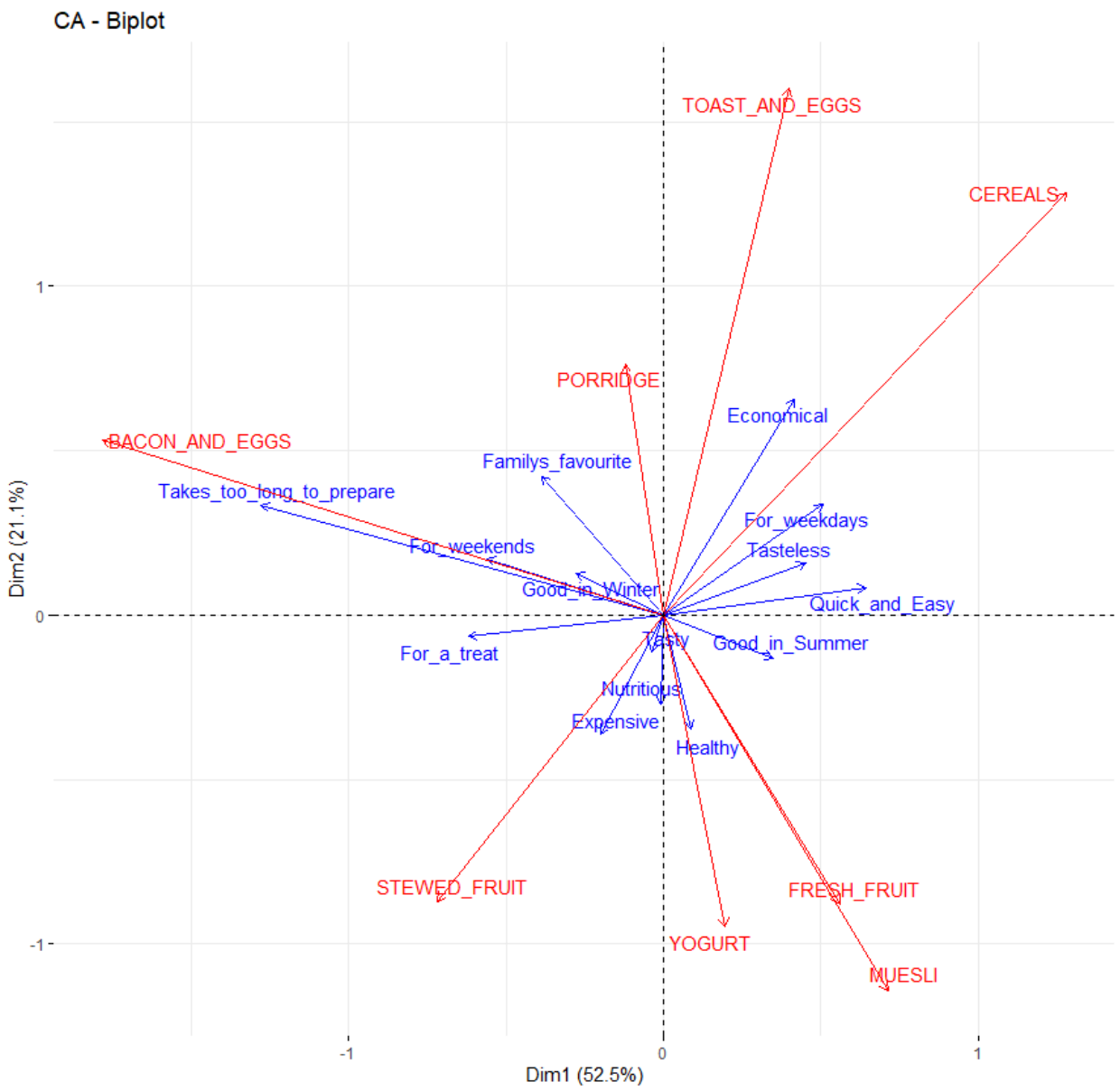



```
> col$contrib
                  Dim 1     Dim 2
CEREALS        25.5299889 25.655570
MUESLI          6.6346171 17.075055
PORRIDGE        0.1599413  6.273402
BACON_AND_EGGS 56.3491874  5.004979
TOAST_AND_EGGS  1.0132285 16.359619
FRESH_FRUIT     3.8879887  9.559437
STEWED_FRUIT    5.9420375  8.748541
YOGURT          0.4830107 11.323396
```

Furthermore, for dimension1, BACON_AND_EGGS contributes the most and then CEREALS follow. In dimension2, CEREALS is the one with the largest contribution. The result from col$contrib is reflected on the bar chart. BACON_AND_EGGS and CEREALS are the only elements that go over expected average value shown as red dot line.

Asymmetric plot (rows and columns)

From this plot, we can take a close look at each arrow with its angle and direction. The more acute the angle between two arrows, the stronger association there will be among rows/columns points. Also, the closer the arrow is to the axis, the higher contribution the rows/columns points will make.



For example, BACON_AND_EGGS has very strong association with For_weekends and then Takes_too_long_to_prepare, Good_in_Winter. Similarly, YOGURT has been highly associated with Healthy and Nutritious.

On the one hand, the arrow toward CEREALS is located in between two axis. It means CEREALS contributes almost equally to each of dimension as col$contrib method already proved.

Determination of the number of axes

Under the consumption of completely random data, each axis has an eigenvalue for rows and columns. In our dataset, the average axis in terms of 14 rows is 14.29% and one in terms of 8 columns is 7.69%. The larger value can be used as a cut-off point to figure out how many axes should be kept. As we set 14.29% to be cut-off point, we can keep first and second axes. First two axes are explaining approximately 52.5% and 21.1% of the total inertia and it is large enough to retain. Meanwhile, the third axe has 11.9% of explanation power. Since it is smaller than the average eigenvalue, it is not enough to stay for further analysis.