DIY Correspondence Analysis

Discovering and sharing stories in data

September 2018

RESEARCHSOFTWARE
A DIVISION OF DISPLAYR

DISPLAYR

Table of Contents

GOAL, OVERVIEW, AND EXAMPLES

This eBook will show you when and how to use correspondence analysis to discover and share stories in data.

The eBook is split into sections describing:

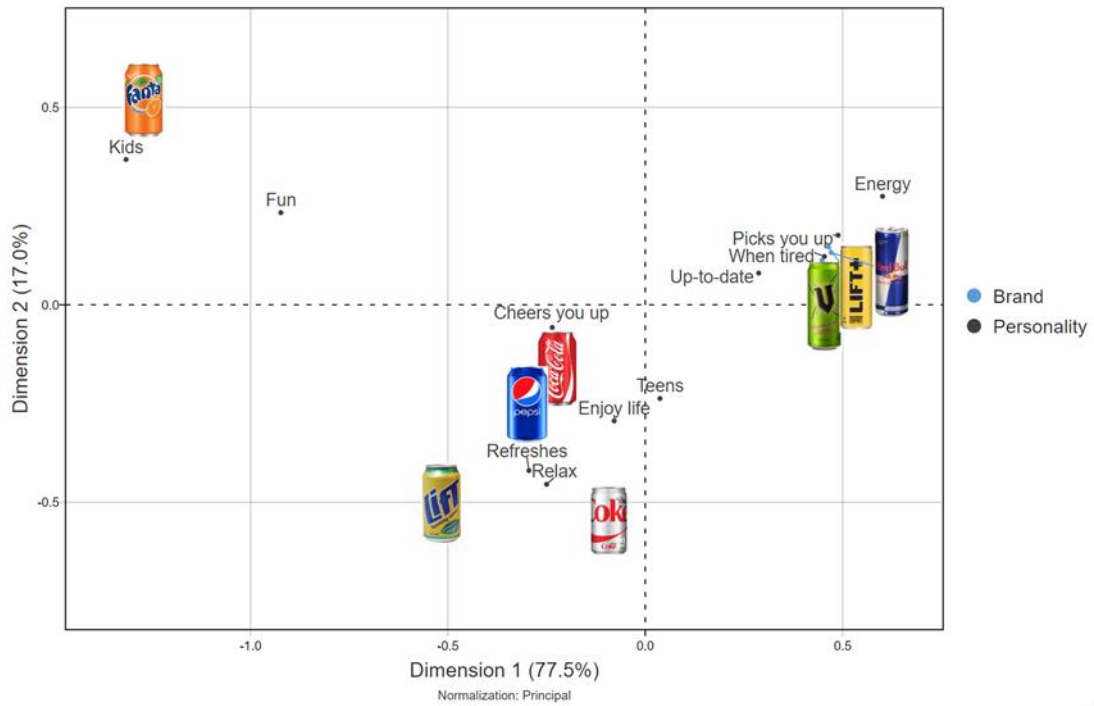
- How correspondence analysis works.
- Which variant of correspondence analysis should be applied to which data.
- Interpretation.
- Visualization.
- Advanced topics (supplementary data points and rotation).

The rest of this chapter presents two examples.

Correspondence analysis is one of the more magical techniques in data analysis. It summarizes the patterns in a table of data as a visualization. Consider the table below. What can you see? Nothing jumps out. Effort is required to find a pattern.

	Coke	V	Red Bull	Lift Plus	Diet Coke	Fanta	Lift	Pepsi
Kids	30%	2%	1%	1%	2%	45%	7%	8%
Teens	69%	46%	41%	24%	18%	13%	11%	22%
Enjoy life	50%	22%	19%	9%	7%	8%	6%	10%
Picks you up	29%	52%	45%	27%	3%	3%	2%	5%
Refreshes	28%	12%	7%	5%	4%	7%	12%	5%
Cheers you up	26%	12%	11%	6%	3%	11%	4%	4%
Energy	19%	55%	47%	28%	1%	2%	2%	3%
Up-to-date	28%	30%	29%	17%	3%	5%	2%	6%
Fun	35%	6%	5%	3%	2%	32%	3%	5%
When tired	25%	38%	33%	19%	3%	2%	1%	4%
Relax	21%	6%	4%	2%	2%	3%	4%	3%

Now look at the visualization below. Patterns not immediately obvious in the table are hard to miss in the visualization.

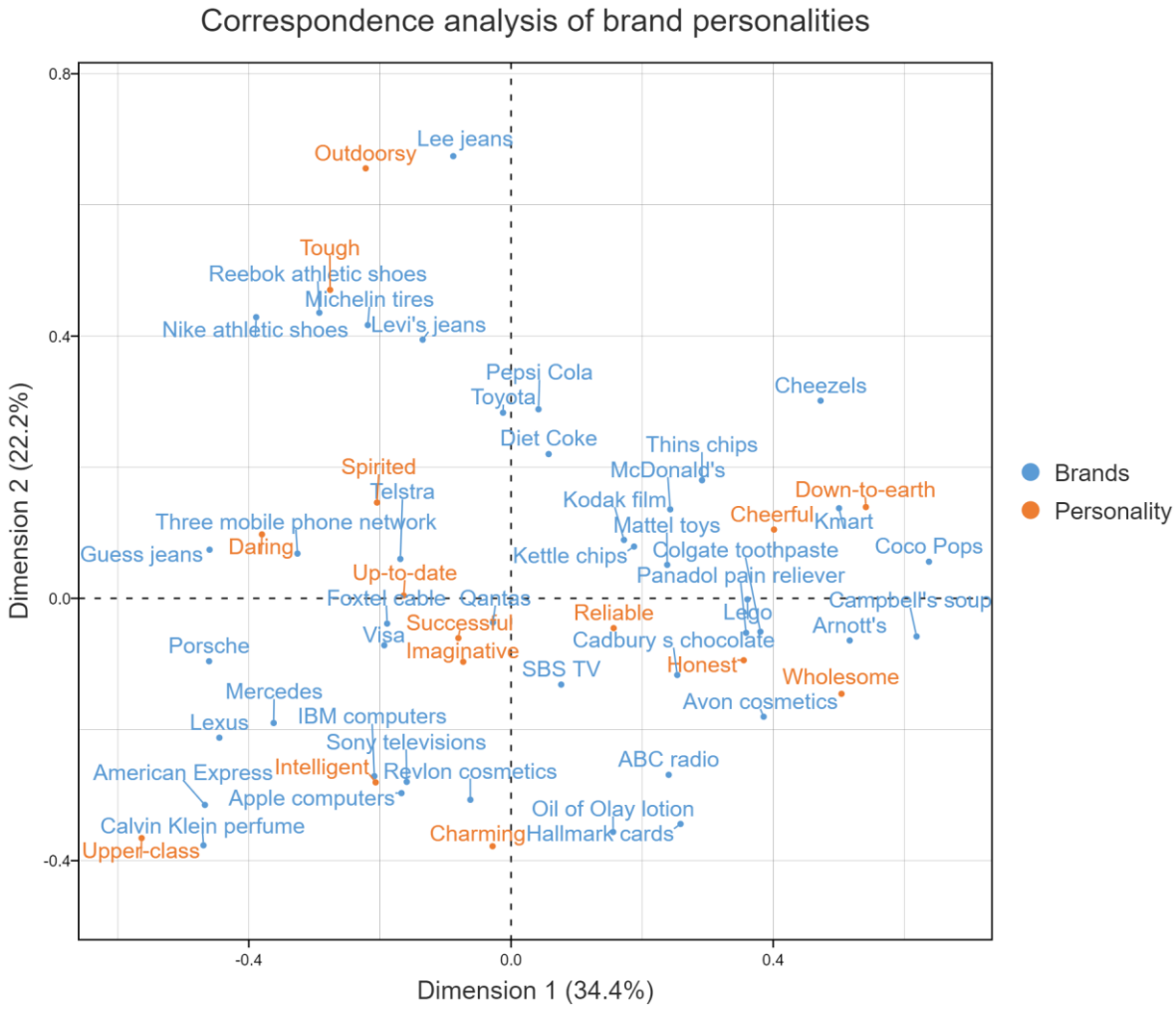


In the example above it is practical to read through the table and find patterns. This next example is much bigger. There is just too much information – 28 brands and 15 attributes – for the human brain to have much of a chance at scanning the table and finding patterns.

	Brand Personality Facets														
Brand associations (%)	**Charming**	**Cheerful**	**Daring**	**Down-to-earth**	**Honest**	**Imaginative**	**Intelligent**	**Outdoorsy**	**Reliable**	**Spirited**	**Successful**	**Tough**	**Up-to-date**	**Upper-class**	**Wholesome**
American Express	20	9	15	10	14	17	44	14	26	14	63	20	30	69	8
Apple computers	25	27	23	14	27	54	71	4	32	34	43	13	53	33	24
Avon cosmetics	33	33	7	26	25	20	10	6	23	7	23	6	20	6	26
Calvin Klein perfume	59	25	50	7	13	31	16	11	21	38	44	5	30	81	10
Campbell's soup	22	30	8	62	50	11	11	16	53	14	31	12	19	11	80
Colgate toothpaste	20	45	13	51	56	17	32	12	82	19	56	23	37	9	52
Diet Coke	13	43	20	21	20	26	9	35	31	29	48	10	43	6	17
Guess jeans	20	18	36	8	6	14	7	22	9	18	21	18	27	36	6
Hallmark cards	58	57	6	28	51	54	24	4	50	24	44	5	23	35	49
IBM computers	11	10	16	12	26	51	75	2	56	18	65	31	55	38	26
Kmart	12	48	12	68	45	17	17	30	46	13	45	12	29	3	45
Kodak film	19	42	8	35	50	33	38	54	67	16	50	12	33	17	29
Lee jeans	12	22	24	40	14	12	5	59	19	23	17	44	24	11	10
Lego	9	51	9	45	49	58	32	6	41	17	37	32	20	5	39
Levi's jeans	27	35	49	47	31	23	16	72	52	47	54	71	46	30	26
Lexus	40	10	32	7	23	29	49	32	42	24	50	25	48	76	13
Mattel toys	15	62	7	24	28	62	18	14	24	17	29	22	30	6	23
McDonald's	10	64	10	33	15	25	15	18	30	16	59	14	39	4	16
Mercedes	52	23	39	6	35	28	61	36	66	35	79	43	55	94	27
Michelin tires	7	12	23	31	38	14	41	66	55	31	40	79	29	33	17
Nike athletic shoes	9	24	48	17	12	43	21	80	38	54	50	58	54	40	16
Oil of Olay lotion	46	20	11	29	33	16	21	11	37	14	28	4	19	39	43
Pepsi Cola	11	53	29	24	18	33	6	37	29	36	37	11	38	7	13
Porsche	50	24	76	5	35	49	64	48	52	64	83	36	54	89	12
Reebok athletic shoes	13	31	44	21	22	41	27	87	41	55	50	54	51	30	17
Revlon cosmetics	45	27	18	19	11	30	13	9	31	14	42	6	32	38	25
Sony televisions	22	24	27	18	35	59	69	5	64	30	61	12	78	41	21
Toyota	19	26	19	41	39	33	41	64	66	36	56	59	48	15	36
Visa	18	18	24	20	24	18	44	22	61	26	69	28	52	40	18

While the plot below is a bit messy, it is much, much, easier to read than the table. At the bottom-left, we can see that that Calvin Klein, American Express, Apple, and Lexus are Upper-class. Porsche mixes Upper-class and Daring. At the top-left, we can see that Tough is shared by Nike, Reebok, Levi's and Michelin, which also are a bit Outdoorsy.

One key tip if you are new to correspondence analysis: the closer anything is to the middle of the map, the less distinct it is. Thus, on this map, we can see that Qantas is poorly described by any of the personality attributes. Similarly, Successful and Imaginative are personality attributes that are not good differentiators between the brands. We can also see that a continuum of sorts is evident in the data. It goes from Upper-class and Intelligent at the bottom-left, through to Cheerful and Down-to-earth at the top-right.



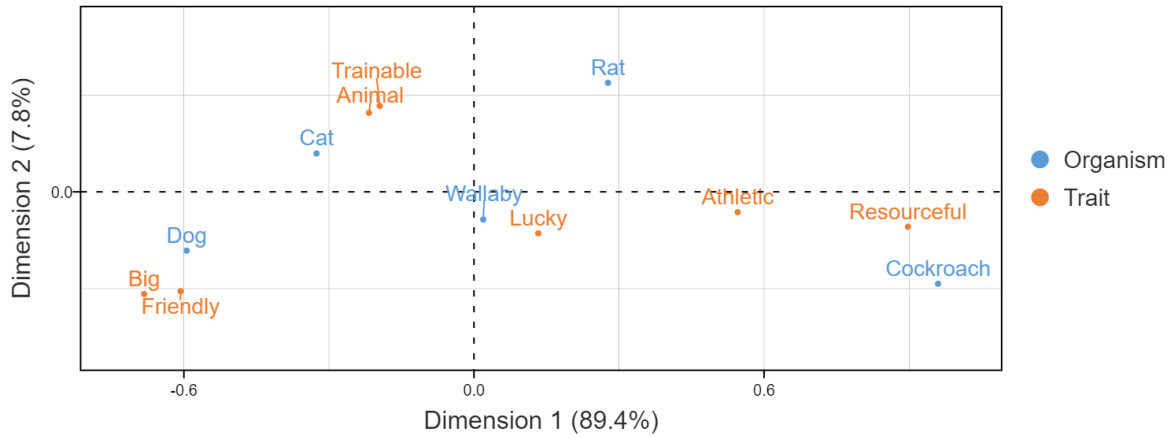
HOW CORREPONDENCE ANALYSIS WORKS

This chapter explains the basics of how correspondence analysis works. It provides a simple worked example, focusing on interpretation.

Much more detail about interpretation is provided in **The correct interpretation of correspondence analysis maps**, and a more mathematically-oriented description in **The math of correspondence analysis**.

The table below shows some data on the traits of some animals, with the resulting correspondence analysis map below. This chapter explains, in simple terms, how the map is computed from the table.

	Big	Athletic	Friendly	Trainable	Resour-ceful	Animal	Lucky
Dog	80	20	90	90	5	100	40
Cat	50	40	40	70	10	100	40
Rat	10	70	20	90	80	99	40
Cockroach	0	80	2	20	95	20	40
Wallaby	35	52	38	47	48	80	40



Step 1: Compute row and column averages

In the first step, compute the averages for each row and column, as shown below.

	Big	Athletic	Friendly	Trainable	Resour-ceful	Animal	Lucky	Average
Dog	80	20	90	90	5	100	40	61
Cat	50	40	40	70	10	100	40	50
Rat	10	70	20	90	80	99	40	58
Cockroach	0	80	2	20	95	20	40	37
Wallaby	35	52	38	47	48	80	40	49
Average	35	52	38	63	48	80	40	51

Step 2: Compute the expected values

Next, for each cell, compute what are known in the trade as the *expected values*. Each cell's expected value is the row average for that cell, multiplied by the column average, and divided by the overall average. So, looking at the averages for Big and Dog in the table on the previous page, we have 35 * 61 / 51 = 42. The following table shows all the expected values.

	Big	Athletic	Friendly	Trainable	Resourceful	Animal	Lucky
Dog	42	63	45	76	57	95	48
Cat	34	51	37	62	47	78	39
Rat	40	60	44	73	55	92	46
Cockroach	25	38	27	46	34	58	29
Wallaby	33	50	36	61	45	76	38

Step 3: Compute the residuals

The residuals are computed by subtracting the expected values from the original data. Thus, for Dog and Big, the residual is 80 – 42 = 38. The residuals are shown below. These residuals are at the heart of correspondence analysis, so do not skip to the next step until you are sure you understand their meaning.

The residuals show the associations between the row and column labels. Big positive numbers indicate a strong positive relationship. The opposite is true for negatives. Let us look at the residuals for Dog. We can see that its biggest score is for Friendly. We can also see that its lowest score is for Resourceful. If you look at the original data table at the beginning of this chapter, neither of these conclusions should surprise you.

The interesting result in the first row is Animal which, for Dog, sits at 100. But, the residual is only 5, indicating virtually no association between being an animal and being a dog. Why? All rows of the data are animals (and four, like the Dog, mammals). So, while a Dog is an animal, this does not

distinguish it from all the other things in the analysis. As a result, this association becomes very weak, which is what is reflected in the residuals.

	Big	Athletic	Friendly	Trainable	Resourceful	Animal	Lucky
Dog	38	-43	45	14	-52	5	-8
Cat	16	-11	3	8	-37	22	1
Rat	-30	10	-24	17	25	7	-6
Cockroach	-25	42	-25	-26	61	-38	11
Wallaby	2	2	2	-14	3	4	2

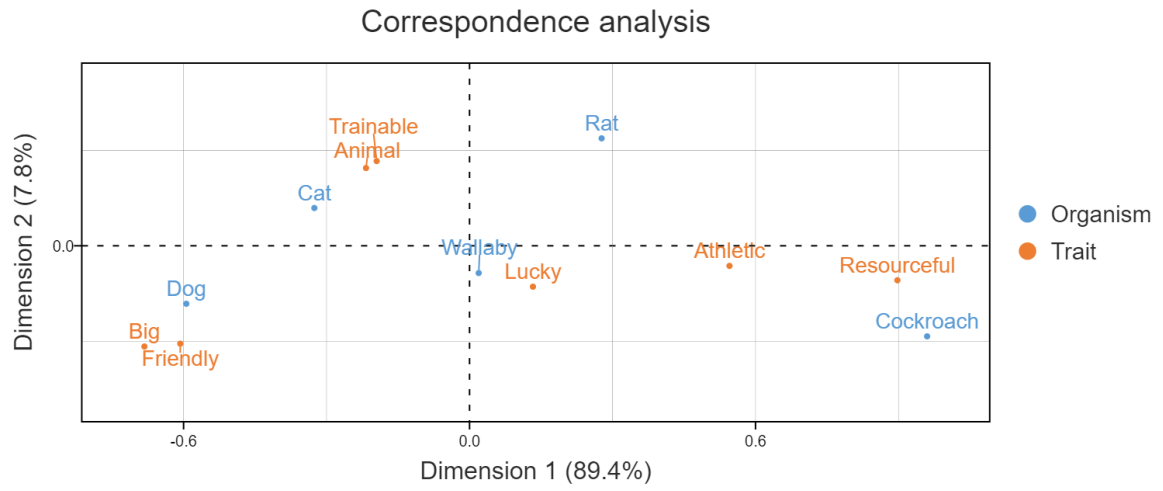
Step 4: Plotting labels with similar residuals close together

Compare the residuals for Cat with those for Dog. While the Dog residuals are generally larger, most are in the same direction. If you take the time, you will realize that in terms of residuals, Dog and Cat are most similar. The next most similar is Dog and Wallaby. Then comes Rat. Last, the Cockroach is least like the Dog. Now look at the blue labels in the plot below, which represent the rows of the table. The relative position of the other animals from Dog in the visualization is consistent with the similarities of their respective residuals.

Now look at the variance shown in the axes labels of the chart. The horizontal dimension explains 89% of the variance in the data whereas the vertical dimension explains only 8%. You can infer the relative amount explained by each dimension on a correctly-drawn map. That is, we can see on this map that the points vary much more on the horizontal than on the vertical, and this is why the relative variance explained of the dimensions varies so greatly.

Together, these two dimensions explain 97% of the variance. This, in turn, tells us that the map represents almost all the information in the residuals, which is good news. If, instead, they only explained a relatively small amount, the map would not tell us the complete story.

Now look at how the columns of the table are represented in the visualization. Big and Friendly are almost equally large, which is why they are next to each other on the map. The least similar trait to Big is Resourceful, which is why it is on the other side of the map to Big.



Step 5: Interpreting the relationship between row and column labels

Now we come to the tricky bit. Correspondence analysis places the row labels on the plot such that the closer two rows (animals) are to each other, the more similar their residuals. This also applies to the column (trait) labels. Most people instinctively conclude that the greater the proximity between a row label and a column label, the higher the residual and association. Wrong. If you think about it for a bit, then you may realize that it is impossible to create a map with such an interpretation (and good careers have been tarnished in the effort to do it.)

To better understand this, compare Dog and Big with Wallaby and Lucky. Dog and Big are close together. Lucky and Wallaby are almost identically proximate. Recall also that the residual for Dog and Big is very high, at 38. Because of this, as we might expect, they are close together on the map. Nevertheless, the residual for Wallaby and Lucky is only 2, yet they are even closer together on the map than Dog and Big. What is going on here?

Now, look at Cockroach. Its residual for Athletic is high at 42. As this is bigger than the 38 for Dog and Big, intuitively you would want Cockroach and Athletic to be very close together on the map. But, Cockroach has an even bigger residual of 61 for Resourceful, and if we put Cockroach and Athletic

next to each other, where can we put Resourceful? There is, in fact, no way to position the labels to sensibly communicate these residuals.

Fortunately, all is not lost. The way that correspondence analysis works means that we can compare between row labels based on distances. We can also compare between column labels based on distances. However, if we want to compare a row label to a column label, we need to:

1. Look at the length of the line connecting the row label to the origin. Longer lines indicate that the row label is highly associated with some of the column labels (i.e., it has at least one high residual).
2. Look at the length of the label connecting the column label to the origin. Longer lines again indicate a high association between the column label and one or more row labels.
3. Look at the angle formed between these two lines. Really small angles indicate association. 90 degree angles indicate no relationship. Angles near 180 degrees indicate negative associations.

Let us work through these rules using some examples. Look at Wallaby and Lucky to the right. The angle is about 30 degrees, indicating some form of association. The short lines, however, suggest that there is either no association, or a very weak one.

A section of the plot showing Cockroach and Athletic is reproduced to the left. The angle is very small, suggesting an association. The arrows are both relatively long, suggesting a strong association. As the arrow to Resourceful would be even longer, and the angle marginally smaller, this tells us that Cockroach is even more strongly associated with Resourceful than with Athletic.

A more in-depth discussion of interpretation can be found in **The correct interpretation of correspondence analysis maps**.

WHICH CORREPONDENCE ANALYSIS TECHNIQUE TO USE WHEN

Four main variants of correspondence analysis are in widespread use. They are all interpreted in the same way, and their difference relates to the type of data that they analyze. The main variants are:

- (Traditional) correspondence analysis
- Correspondence analysis of square tables
- Multiple correspondence analysis
- Correspondence analysis of multiple tables

This section of the eBook describes when to apply each of these techniques, as well as when not to apply correspondence analysis.

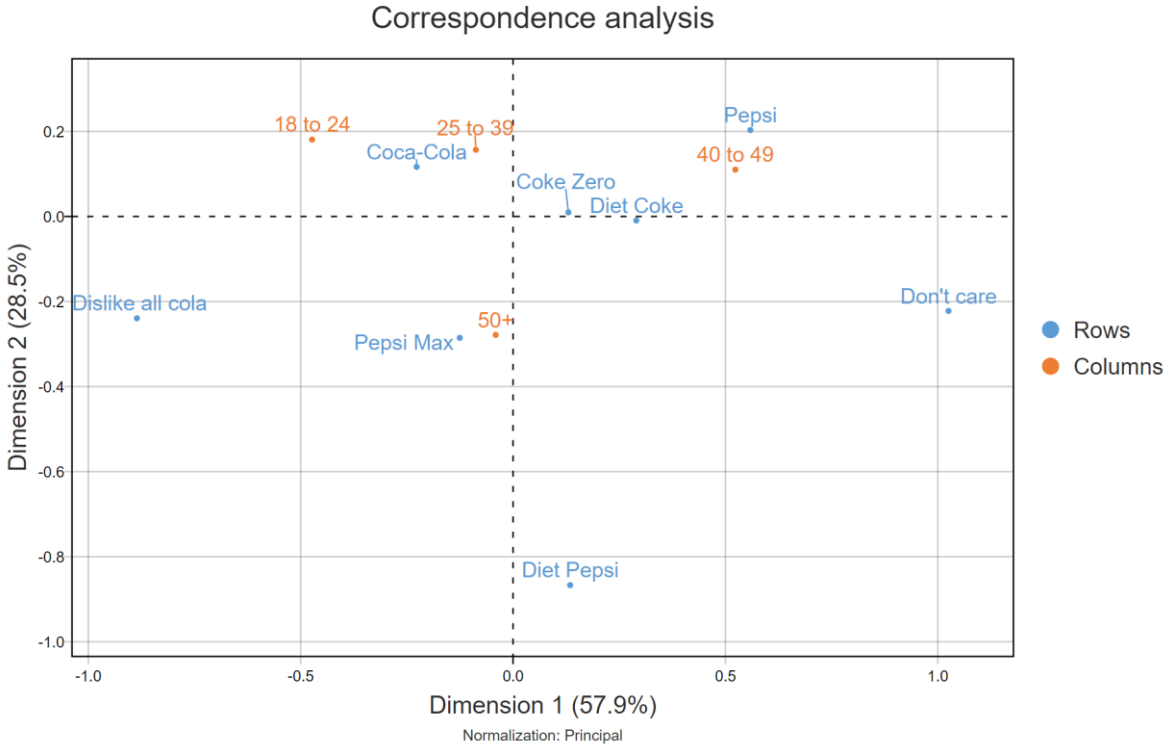
(Traditional) correspondence analysis

> The term *correspondence analysis* refers both to a specific technique, which is what has been discussed in the chapter so far, and a set of closely related techniques. When there is a chance of misunderstanding about the difference between the two usages, this eBook refers to the technique as *traditional correspondence analysis.*
>
> This chapter discusses applications of traditional correspondence analysis. Other variants of correspondence analysis are discussed in the following chapters.

The classic application for correspondence analysis is the analysis of *contingency tables*. A contingency table is a table where the row categories are mutually exclusive, the column categories are also mutually exclusive, and the cells show the number of observations in each category. For example, the contingency table below show the number of people who prefer different brands of cola by age. This data is suitable for analysis using correspondence analysis.

Count	18 to 24	25 to 39	40 to 49	50 to 64	65 or more
Coca-Cola	63	131	44	84	19
Diet Coke	2	33	24	18	12
Coke Zero	10	53	32	40	8
Pepsi	6	18	32	13	3
Diet Pepsi	0	2	3	10	5
Pepsi Max	15	31	15	49	9
Dislike all cola	3	0	0	3	0
Don't care	0	0	6	4	0

The resulting correspondence analysis visualization is shown below. This is often referred to as a *market map* in research, or sometimes just a *map*. The essence of interpretation is that: (1) the further any label is from the origin of the map (where the dashed lines cross, 0,0), the more the map has to say about that label; (2) where two labels are in the same direction, they are related. Thus, this map reveals that Diet Pepsi is more preferred by people aged 50+, Coca-Cola by people aged 18 to 24 and Pepsi by people aged 40 to 49. More detail is provided about interpretation in the rest of this chapter and in the section on **INTERPRETATION**.



The table below is almost a contingency table - “almost” because it includes the row and column totals (labeled as NET). It is not valid to analyze such a table using correspondence analysis, unless you exclude the NET or total columns (this occurs automatically in Q and Displayr).

Count	18 to 24	25 to 39	40 to 49	50 to 64	65 or more	NET
Coca-Cola	63	131	44	84	19	341
Diet Coke	2	33	24	18	12	89
Coke Zero	10	53	32	40	8	143
Pepsi	6	18	32	13	3	72
Diet Pepsi	0	2	3	10	5	20
Pepsi Max	15	31	15	49	9	119
Dislike all cola	3	0	0	3	0	6
Don't care	0	0	6	4	0	10
NET	99	268	156	221	56	800

Percentages and index values

The mathematics used to analyze correspondence analysis is derived from the assumption that the data is a contingency table. Nevertheless, the technique works effectively for many other types of data, provided that the data is all on the same scale. For example, it is usually useful to apply correspondence analysis to tables showing row percentages, column percentages, and index values. However, each will give a different output, as each analysis emphasizes different aspects of the data, and these aspects are emphasized by the resulting correspondence analyses.

Often, it is preferable to analyze the data using percentages, means, or index values rather than tables showing *counts,* such as the one used in the previous section. Typically, the most useful correspondence analysis will be the one where the input table is the most informative. For example:

- Where there is missing data, percentages are often better than counts (this is particularly the case with multiple response tables, shown in the next section), as the counts will often be biased.

- With weighted samples, means, percentages, and index values are often be better than counts.
- When certain groups in the population are over or under-represented, percentages will often be better than counts.

Multiple response tables

Most tables that show multiple response data can also be used with correspondence analysis. The table below, which is referred to as a *brand association grid* by market researchers, is made up of the data from 63 different variables. Each of the 800 respondents in the data set has indicated which brands possess which attributes. As the data is non-negative, and is all on the same scale, it is a prime candidate for correspondence analysis (but we would usually want to exclude the NET row and column, as is automatically done by Q and Displayr).

%	Feminine	Health-conscious	Innocent	Older	Open to new experiences	Rebellious	Sleepy	Traditional	Weight-conscious	NET
Coke	6%	2%	11%	65%	22%	26%	10%	91%	1%	98%
Diet Coke	57%	58%	22%	23%	9%	5%	23%	15%	76%	92%
Coke Zero	22%	54%	11%	5%	51%	64%	10%	3%	64%	95%
Pepsi	9%	3%	10%	39%	17%	18%	14%	55%	0%	80%
Diet Pepsi	62%	58%	45%	10%	17%	4%	30%	4%	77%	95%
Pepsi Max	9%	31%	7%	7%	49%	45%	6%	4%	40%	86%
None of these	9%	17%	30%	7%	13%	15%	39%	3%	6%	58%
NET	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tables of means

The table below shows averages. It meets all the requirements for correspondence analysis. (However, as there are only two rows of data once the SUMs are excluded, the resulting map will show all the data points organized along a straight line, which can cause a bit of a panic if you are not expecting it.)

Average	Coca-Cola	Diet Coke	Coke Zero	Pepsi	Diet Pepsi	Pepsi Max	SUM
'out and about'	1.8	.8	2.0	.4	.3	.9	6.2
'at home'	3.6	2.3	3.7	.5	.1	1.6	11.9
SUM	5.4	3.1	5.7	.9	.4	2.5	18.0

Raw numeric data

The table below shows the raw data. Each row represents a person. Each column indicates the consumption of the different brands either at home or out and about. Is this data suitable for correspondence analysis? Raw data is usually OK, provided that:

- The data is either binary or numeric. Raw data for unordered categorical variables (e.g., occupation, brand preference), will not work, as the data has no meaningful scaling (i.e., averages do not make sense): the analysis of such raw data can be addressed with **Multiple correspondence analysis**.
- We have no missing data in the rows. If we do have missing data, the best solution is usually to use imputation to replace the missing values, although sometimes it can be OK to filter the data so that the rows containing missing data are removed.

Values	Coke, Out	Coke, Home	Diet Coke, Out	Diet Coke, Home	Zero, Out	Zero, Home	Pepsi, Out	Pepsi, Home
1	.00	.00	1.00	1.00	.00	.00	.00	.00
2	.00	.00	.00	.00	.00	1.00	1.00	.00
3	.00	.00	9.00	16.00	.00	.00	.00	.00
4	.00	.00	.00	.00	.00	.00	.00	.00
5	2.00	8.00	1.00	1.00	.00	1.00	.00	.00
6	.00	.00	.00	.00	.00	.00	.00	.00
7	2.00	8.00	.00	.00	.00	.00	.00	1.00
8	.00	.00	.00	.00	.00	3.00	.00	.00
9	8.00	2.00	.00	3.00	.00	3.00	.00	.00
10	.00	2.00	.00	.00	.00	5.00	.00	.00
11	.00	.00	.00	2.00	4.00	24.00	.00	.00
12	1.00	2.00	.00	1.00	.00	1.00	.00	.00
13	.00	.00	1.00	1.00	9.00	20.00	.00	.00
14	2.00	.00	.00	.00	.00	.00	.00	1.00
15	4.00	14.00	.00	.00	.00	.00	.00	.00
16	.00	.00	3.00	15.00	1.00	.00	.00	.00
17	.00	.00	.00	.00	.00	.00	.00	.00
18	.00	2.00	.00	.00	.00	2.00	.00	.00
19	.00	.00	.00	.00	.00	.00	.00	1.00
20	.00	2.00	.00	.00	.00	.00	.00	.00

The cool thing about using raw data is we can understand the distribution of respondents in the data. In the visualization below, each orange dot represents a row of data (i.e., a person). It shows us that Coke, Out and Coke, Home, have lots of people associated with them (indicating high levels of consumption). The hard thing is that all the usual rules of interpretation apply (these are discussed in detail in **INTERPRETATION**).

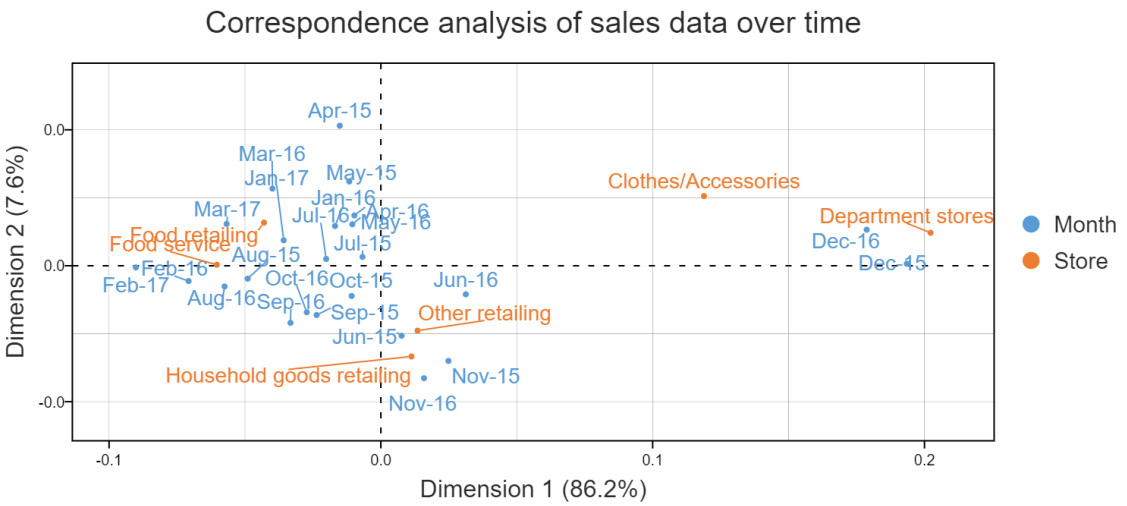

Diet Coke, Out
Diet Coke, Home
Coke, Out
Coke, Home
Diet Pepsi, Home
Zero, Home
Zero, Out
Diet Pepsi, Out
Pepsi, Out
Pepsi, Home
Max, Out
Max, Home
1.2
0.0
-1.2
-2.4
-1.0
0.0
1.0
Dimension 2 (19.4%)
Dimension 1 (24.1%)

Time series data (e.g., sales)

The final example, shown below, shows sales by different retailer categories by month. While you may not think of sales data as being appropriate for correspondence analysis, it satisfies all the criteria.

	Food retailing	Household goods retailing	Clothes/ Accessories	Department stores	Other retailing	Food service
Apr-15	9538	3714	1793	1442	3059	3267
May-15	9707	3977	1891	1429	3215	3314
Jun-15	9218	4319	1775	1481	3244	3258
Jul-15	9718	4186	1789	1541	3337	3445
Aug-15	9835	4160	1697	1332	3380	3421
Sep-15	9646	4274	1790	1400	3455	3444
Oct-15	10319	4568	1889	1566	3580	3526
Nov-15	10129	4693	1916	1731	3825	3491
Dec-15	11731	5736	3080	2914	4643	3820
Jan-16	10245	4377	1876	1519	3305	3432
Feb-16	9557	3980	1599	1156	3257	3187
Mar-16	10354	4097	1781	1452	3399	3435
Apr-16	9728	4065	1925	1451	3356	3452
May-16	9815	4093	1927	1450	3429	3431
Jun-16	9517	4357	1967	1596	3414	3314
Jul-16	9929	4225	1876	1468	3493	3573
Aug-16	10042	4239	1806	1294	3562	3648
Sep-16	10006	4469	1897	1394	3602	3696
Oct-16	10483	4697	1938	1497	3643	3717
Nov-16	10436	4874	2057	1684	4051	3679
Dec-16	12230	5782	3331	2850	4860	4047
Jan-17	10432	4464	1940	1429	3422	3621
Feb-17	9575	3898	1559	1092	3231	3261
Mar-17	10510	4197	1827	1370	3590	3619

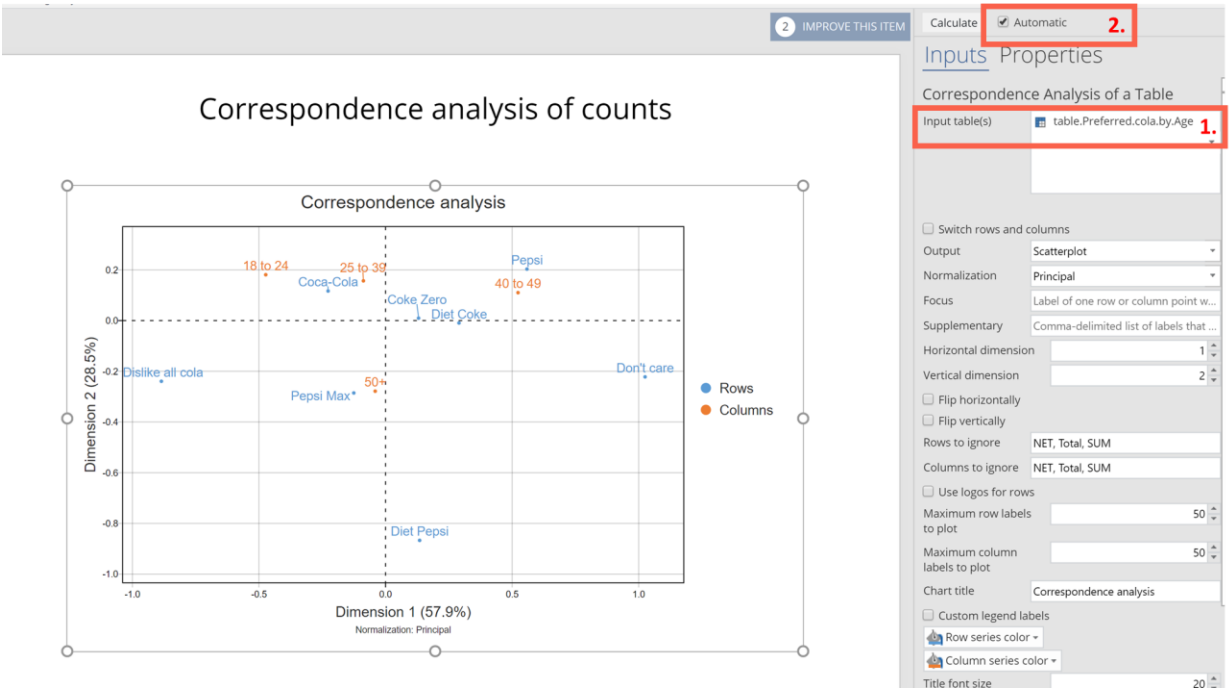
The visualization below displays the sales data. It shows that department store and clothes/accessory sales are strongly associated with December.



Software

All the examples in this eBook have been created using Q (www.q-researchsoftware.com) and Displayr (www.Displayr.com).The process for creating correspondence analysis is:

- Create a table.
- Either:
    - In Q, select **Create > Dimension Reduction > Correspondence Analysis of a Table**.
    - In Displayr, select **Insert > Dimension Reduction > Correspondence Analysis of a Table**.
- Select the table in the **Input table(s)** field (**1.**; see the screenshot below)
- Check **Automatic** (**2.**)

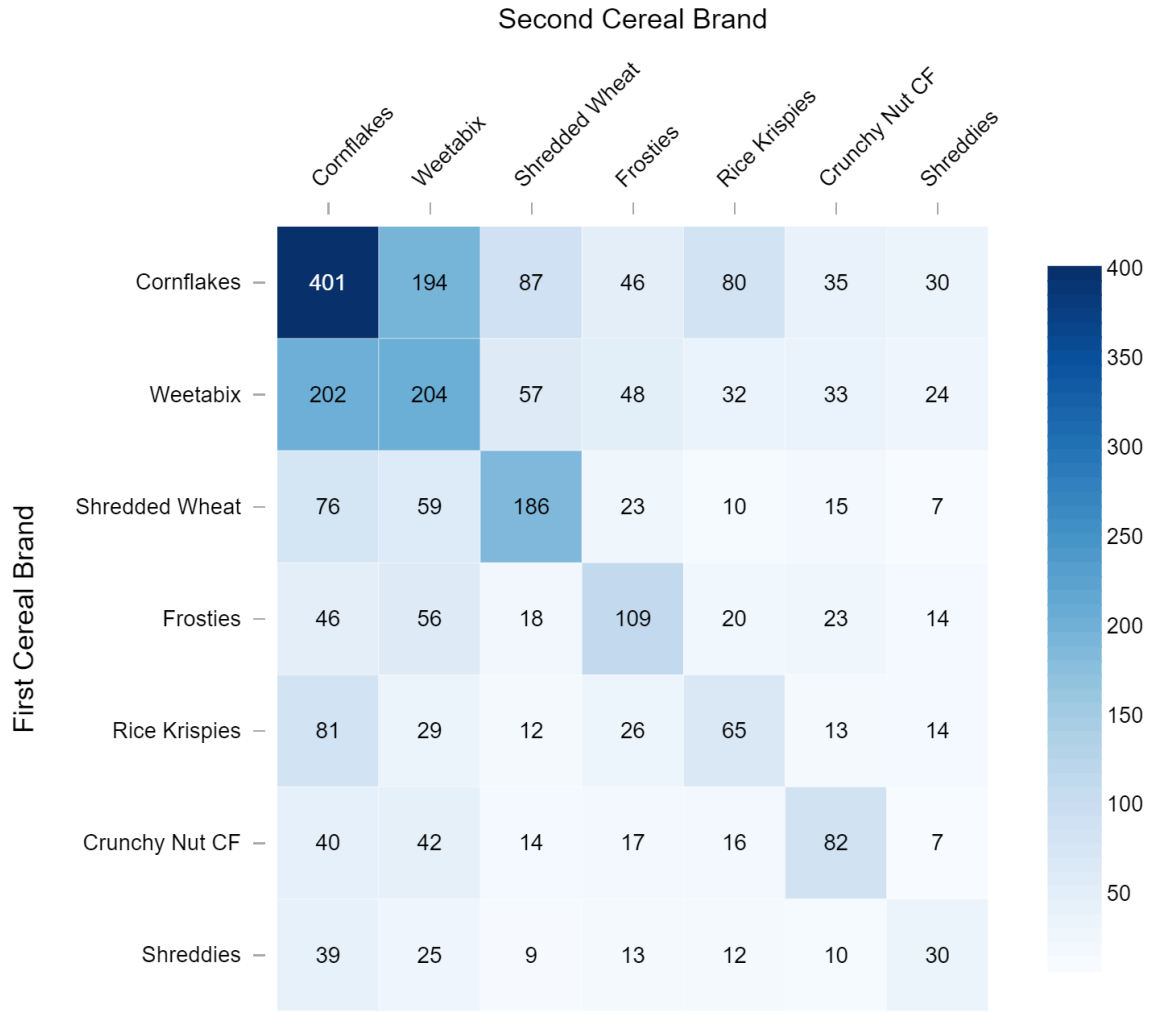

IMPROVE THIS ITEM
Calculate
Automatic
2.
Inputs
Properties
Correspondence Analysis of a Table
Input table(s)
table.Preferred.cola.by.Age
1.
Switch rows and columns
Output
Scatterplot
Normalization
Principal
Focus
Label of one row or column point w...
Supplementary
Comma-delimited list of labels that ...
Horizontal dimension
1
Vertical dimension
2
Flip horizontally
Flip vertically
Rows to ignore
NET, Total, SUM
Columns to ignore
NET, Total, SUM
Use logos for rows
Maximum row labels to plot
50
Maximum column labels to plot
50
Chart title
Correspondence analysis
Custom legend labels
Row series color
Column series color
Title font size
20
Correspondence analysis of counts
Correspondence analysis
18 to 24
25 to 39
Pepsi
Coca-Cola
40 to 49
Coke Zero
Diet Coke
Dislike all cola
Don't care
50+
Pepsi Max
Diet Pepsi
Rows
Columns
Dimension 2 (28.5%)
Dimension 1 (57.9%)
Normalization: Principal

Correspondence analysis of square tables

A *square table* is one where the row and column labels are identical. Common examples of square tables include tables showing duplication data (e.g., which magazines are read by readers of which other magazines) and switching matrices (which brands people switch between), which are also known as transition tables and confusion matrices.

Although square tables can be analyzed using traditional correspondence analysis, it is better to use algorithms specifically designed for this type of data, as the resulting outputs are neater and, more importantly, these analyses allow for an understanding of dynamics.

The table below shows which brands of breakfast cereal people consumed on two subsequent occasions.[1] For example, we can see that 401 purchased Cornflakes twice in a row, 194 purchased Cornflakes and then Weetabix, etc. As the labels in the rows and columns are identical, we can analyze the data using a special variant of correspondence analysis designed for such data.[2]



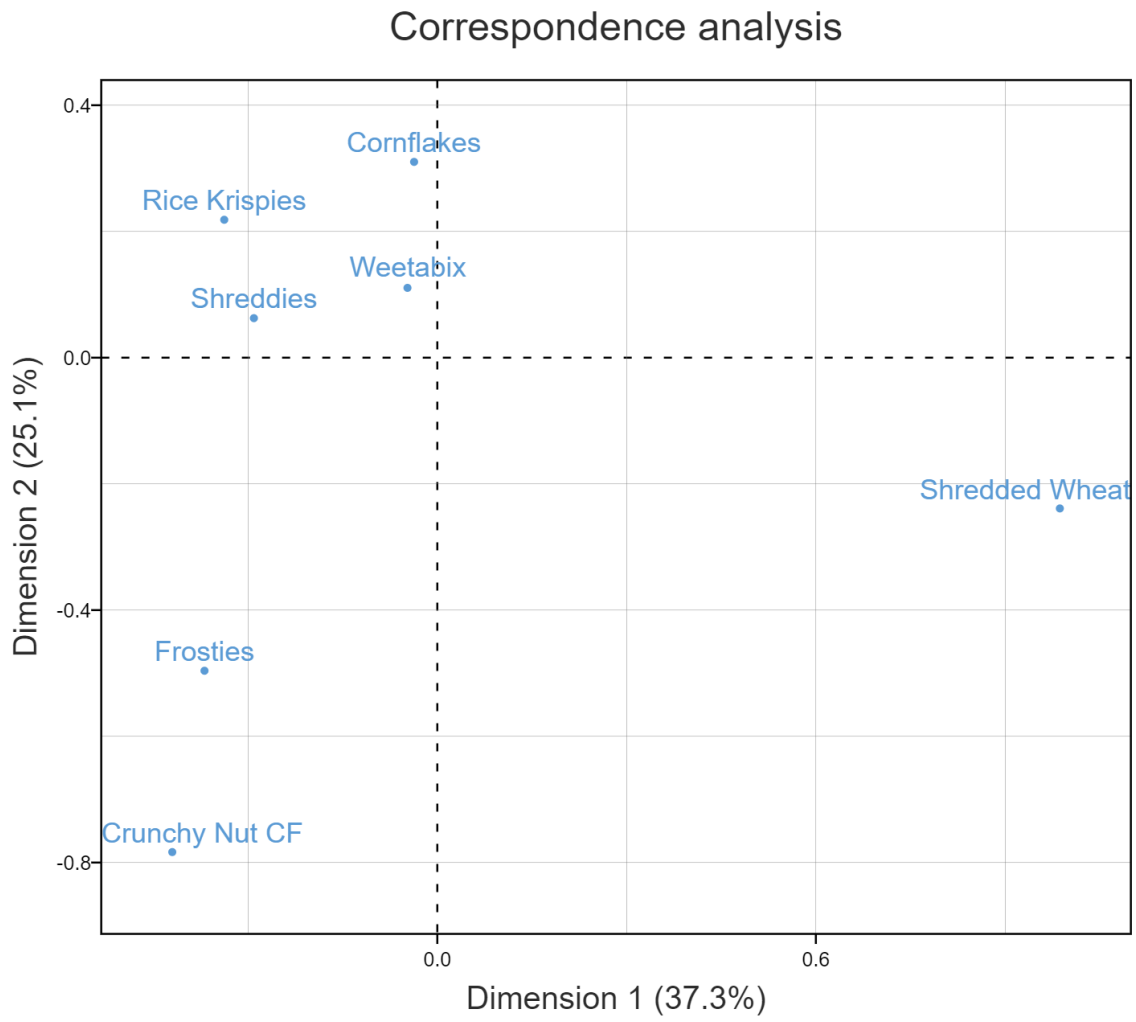
[1] The data is from Dawes, John (2007). "The Structure of Switching: An Examination of Market Structure Across Brands and Brand Variants", *ANZMAC Conference Proceedings*, University of Otago, November.

[2] Greenacre, M. (2000). "Correspondence analysis of square asymmetric matrices." Applied Statistics 49(3): 297-310.

Before delving into the correspondence analysis, let's take a look at the data above. One of the first observations that we can make about it is the strong *main diagonal:* people tend to buy the same cereal repeatedly.

Looking away from the diagonal, there is also high symmetry. For example, the numbers switching from Cornflakes to Rice Krispies (80) is almost the same as switching in the other direction (81). Both of these observations are quite typical of square tables from consumer data.

Now let's perform the correspondence analysis of the square table (using the variant of correspondence analysis designed for such data). The scatterplot below shows the first two output dimensions.



It's tempting to draw immediate conclusions from the plot above. Before we do so, we need to take note of a few things.

Any square matrix can be broken down into *symmetric* and *skew-symmetric components*. The correspondence analysis of those two components is driven by different aspects of the data. The symmetric component shows us how much two-way exchange occurs between categories (i.e., substitution). The skew-symmetric component determines the net flow into or out of a category.

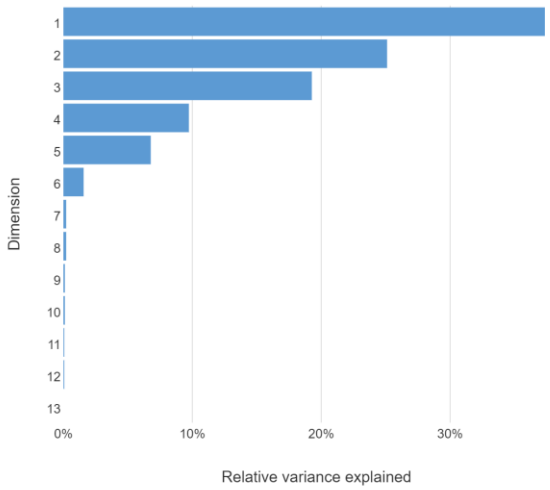
We can tell which dimensions are symmetric and which are skew-symmetric by inspecting how much variance each dimension explains. The table below is produced in Q and Displayr by changing the setting of **Output** to **Text.**

The symmetric component produces dimensions that each explain a different amount of the variation in the table. In more technical language, the canonical correlations are unique. Correspondence analysis of square tables produces dimensions that occur in pairs with the same amount of variation, and these show the skew-symmetric component. Looking at the table below, we can see that dimensions 1 to 6 and 13 all show symmetric components, whereas the 7 and 8, 9 and 10, and 11 and 12, are pairs of skew-symmetric components (as they have the same canonical correlations).

```
Correspondence analysis of a square table

Inertia(s):
             Canonical Correlation      Inertia Proportion explained
Dimension 1           4.221763e-01 1.782328e-01          3.730594e-01
Dimension 2           3.460415e-01 1.197447e-01          2.506379e-01
Dimension 3           3.030568e-01 9.184344e-02          1.922377e-01
Dimension 4           2.151026e-01 4.626915e-02          9.684603e-02
Dimension 5           1.792711e-01 3.213812e-02          6.726836e-02
Dimension 6           8.518357e-02 7.256241e-03          1.518805e-02
Dimension 7           2.784087e-02 7.751141e-04          1.622393e-03
Dimension 8           2.784087e-02 7.751141e-04          1.622393e-03
Dimension 9           1.725646e-02 2.977853e-04          6.232949e-04
Dimension 10          1.725646e-02 2.977853e-04          6.232949e-04
Dimension 11          8.049223e-03 6.478999e-05          1.356120e-04
Dimension 12          8.049223e-03 6.478999e-05          1.356120e-04
Dimension 13          4.941547e-17 2.441889e-33          5.111121e-33
```

If we plot the proportion of variance explained (the last column), we can see that the skew symmetric components (dimensions 7 to 11) are trivial, and can be ignored.



The absence of any serious skew symmetric components has an important implication. It tells us that the data is consistent with a *steady state system*, which is a fancy way of saying that the switching between the brands is routine, perhaps caused by people liking to regularly switch between different sets of brands for variety, or due to purchasing for multiple members of the household.

The table below shows some data from the 1840s. The rows tell us the occupation of German politicians prior to entering parliament, and the columns show their occupations after leaving parliament.[3]

This table shows some clear asymmetry. For example, we can see that only 11 people who were in administration prior to enter parliament left parliament to enter Justice, whereas 83 people who were initially working in justice left to work in administration.

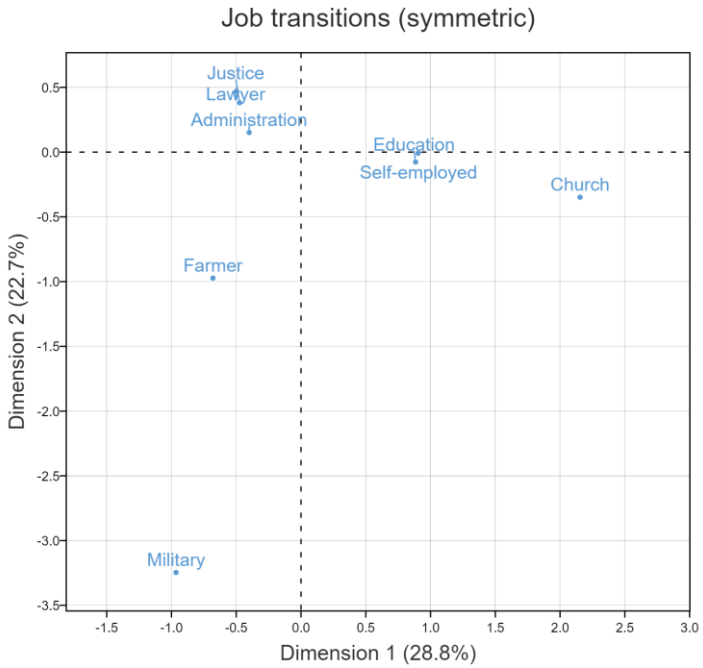
	Justice	Administ-ration	Education	Military	Church	Farmer	Lawyer	Self-employed
Justice	117	83	19	0	1	17	76	6
Administration	11	37	6	0	0	7	7	6
Education	4	3	67	1	4	3	5	13
Military	0	5	1	16	0	8	1	0
Church	0	1	11	0	26	0	1	0
Farmer	0	4	0	0	0	19	1	0
Lawyer	8	2	1	0	0	0	22	1
Self-employed	0	7	20	0	4	1	0	37

[3] The example is from Greenacre, M. (2000). "Correspondence analysis of square asymmetric matrices." *Applied Statistics* 49(3): 297-310.

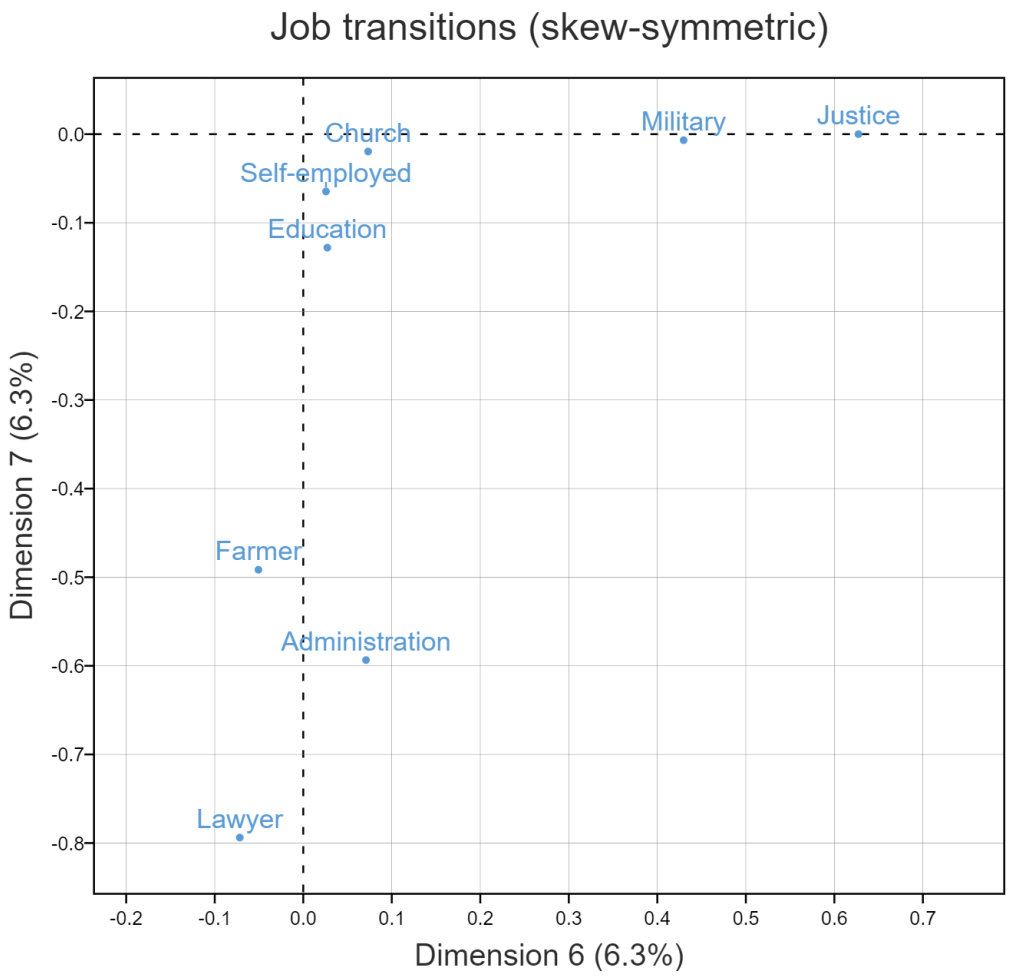
Looking at the proportion of variance explained, we can see the first four dimensions are all symmetric. However, the first pair of asymmetric dimensions, 5 and 6, jointly explain 13% of the variance, which is much more than was the case in the earlier cereal data set, showing that the data is not from a steady state system (i.e., there are dynamic relationships, with category sizes changing over time).

	Canonical Correlation	Inertia	Proportion explained
Dimension 1	0.79996602	0.63994562775	0.28796920567
Dimension 2	0.70975098	0.50374646051	0.22668092694
Dimension 3	0.61702417	0.38071882719	0.17131970826
Dimension 4	0.41031304	0.16835679188	0.07575889189
Dimension 5	0.38389762	0.14737738415	0.06631836582
Dimension 6	0.37309236	0.13919791006	0.06263768335
Dimension 7	0.37309236	0.13919791006	0.06263768335
Dimension 8	0.24834091	0.06167320933	0.02775233447
Dimension 9	0.11929967	0.01423241166	0.00640444454
Dimension 10	0.11929967	0.01423241166	0.00640444454
Dimension 11	0.08168950	0.00667317454	0.00300286257
Dimension 12	0.08168950	0.00667317454	0.00300286257
Dimension 13	0.01103286	0.00012172407	0.00005477463
Dimension 14	0.00787490	0.00006201405	0.00002790571
Dimension 15	0.00787490	0.00006201405	0.00002790571

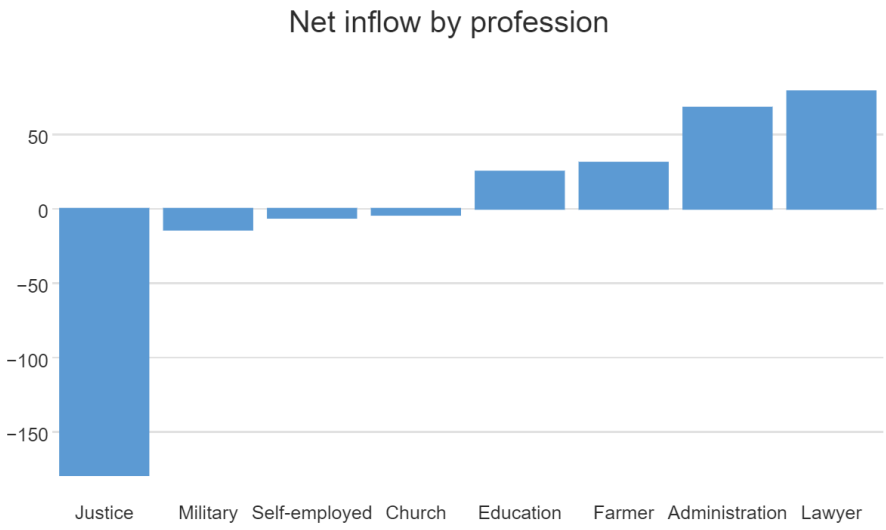
The map for the first two dimensions is shown below, and it reveals two clusters of occupations that involve relatively high levels of switching between them: justice, lawyer, and administration, and, education and self-employment.



Dimensions 5 and 6 are plotted on the next page. Lawyer, justice, and administration are at the extremities. This means that those professions experience a relatively high net inflow and outflow.



We cannot say from the chart which has the inflow, and which has the outflow. The only way to tell is to look at the raw data. To clarify this point, we can compute the net inflow for each professional by working out the difference between the row totals and column totals for each profession in the original table. The final column chart shows us that Lawyer has the inflow, and Justice has the outflow.



Software

In Q and Displayr the process for creating correspondence analysis of a square table is:

- Create a square table.
- Either:
    - In Q, select **Create > Dimension Reduction > Correspondence Analysis of a Square Table**.
    - In Displayr, select **Insert > Dimension Reduction > Correspondence Analysis of a Square Table**.
- Select the table in the **Input table(s)** field.
- Specify which dimension to plot in the **Horizontal dimension** and **Vertical dimension**. By default, these show **1** and **2**.
- Check **Automatic**.

Multiple correspondence analysis

> Multiple correspondence analysis sounds better than correspondence analysis. However, for 99% of real-world data problems, traditional correspondence analysis is the more useful technique.
>
> This chapter explains when multiple correspondence analysis is useful.

When most people do introductory statistics they first learn regression, and then the more useful multiple regression. So, when people come across multiple correspondence analysis they often assume that just like with regression, the "multiple" in correspondence analysis denotes the technique being more useful. Unfortunately, this is not the case. Multiple correspondence analysis has a role, but in the main it is much less useful than traditional correspondence analysis.

For most purposes the way to think about the two techniques is this:

- Correspondence analysis is a technique for summarizing relativities in tables. As tables are ubiquitous in data analysis, it is a technique that can be used widely.
- Multiple correspondence analysis is a technique for analyzing categorical variables. It is essentially a form of factor analysis for categorical data. You should use it when you want a general understanding of how categorical variables are related.

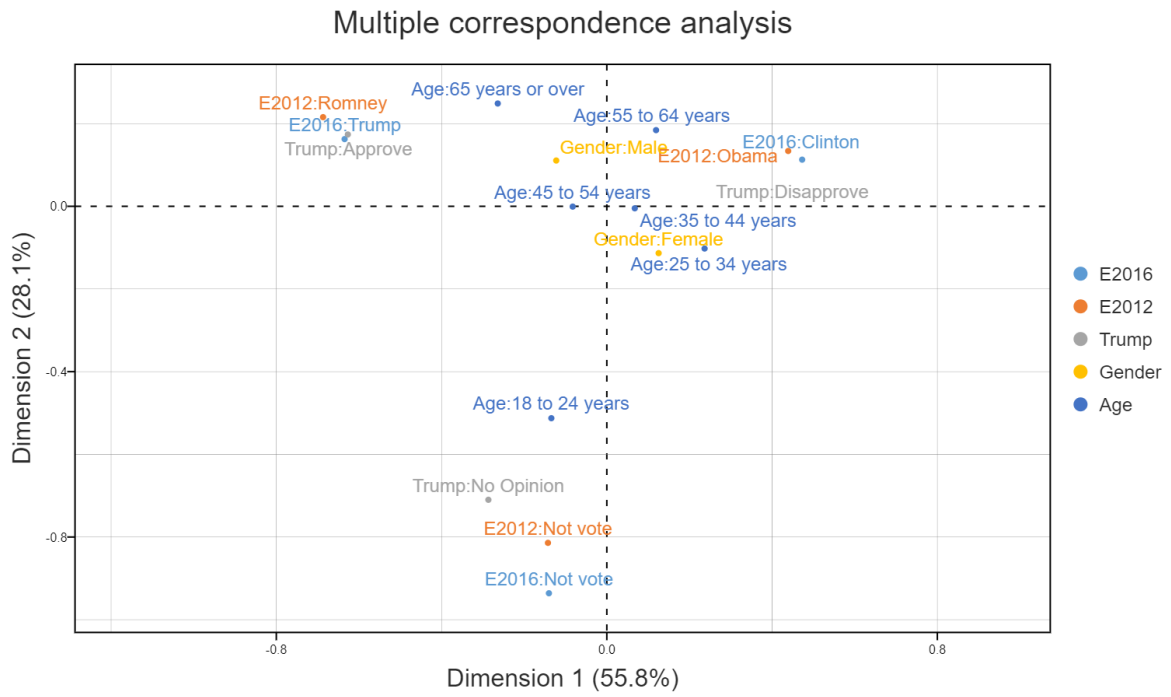
Both techniques give the same answer when you have two variables. You can also use both for more than two variables, but they give different answers, as illustrated below.

The reason for the word "multiple" is that multiple correspondence can be applied to a table that has more than two dimensions (e.g., a cube), whereas correspondence analysis requires as an input a table with only two dimensions. So, the word "multiple" refers to the number of dimensions of the input table. A five-dimensional table is shown later in the chapter.

An example of multiple correspondence analysis

The scatterplot below shows a multiple correspondence analysis of five variables: voting in the 2012 and 2016 US elections, approval of President Trump, age, and gender. The key conclusions from it are that:

- People aged 18 to 24 were less likely to vote and more likely to have no opinion about Trump.
- Approval and disapproval is correlated with candidate-party choice for 2012 and 2008.



The problems with multiple correspondence analysis

Difficulty in checking the input data table

As is discussed in detail in **The correct interpretation of correspondence analysis maps**, it is key to check key conclusions in the data when making conclusions from correspondence analysis. However, this is very difficult to do with multiple correspondence analysis. The above plot seems useful. And multiple correspondence analysis can be useful. Nevertheless, it has some serious limitations. The first limitation is that it is extremely difficult to check conclusions by looking at the raw data. Check out the table below. It is a five-dimensional table and goes over multiple pages. As a result, it would be extremely hard to use it to check your conclusions from the plot. It shows counts, as it is difficult to even think about how to compute percentages on a five-dimensional table.

```
, , Trump = Approve, Gender = Male, Age = 18 to 24 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      3        0
  Not vote     0      0        0

, , Trump = Disapprove, Gender = Male, Age = 18 to 24 years

          E2012
E2016      Obama Romney Not vote
  Clinton      2      0        1
  Trump        0      0        0
  Not vote     0      0        2

, , Trump = No Opinion, Gender = Male, Age = 18 to 24 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      1        1
  Not vote     0      0        1

, , Trump = Approve, Gender = Female, Age = 18 to 24 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      2        0
  Not vote     0      0        1

, , Trump = Disapprove, Gender = Female, Age = 18 to 24 years

          E2012
E2016      Obama Romney Not vote
  Clinton      5      0        3
  Trump        0      0        1
  Not vote     0      0        0

, , Trump = No Opinion, Gender = Female, Age = 18 to 24 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      2        0
  Not vote     0      0        3

, , Trump = Approve, Gender = Male, Age = 25 to 34 years

          E2012
```

```
E2016      Obama Romney Not vote
  Clinton      1      0        0
  Trump        0      4        0
  Not vote     0      1        1

, , Trump = Disapprove, Gender = Male, Age = 25 to 34 years

          E2012
E2016      Obama Romney Not vote
  Clinton      7      0        0
  Trump        1      0        0
  Not vote     0      0        1

, , Trump = No Opinion, Gender = Male, Age = 25 to 34 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      0        0
  Not vote     0      0        0

, , Trump = Approve, Gender = Female, Age = 25 to 34 years

          E2012
E2016      Obama Romney Not vote
  Clinton      1      0        0
  Trump        1      1        2
  Not vote     0      0        0

, , Trump = Disapprove, Gender = Female, Age = 25 to 34 years

          E2012
E2016      Obama Romney Not vote
  Clinton     16      1        0
  Trump        0      0        1
  Not vote     0      0        2

, , Trump = No Opinion, Gender = Female, Age = 25 to 34 years

          E2012
E2016      Obama Romney Not vote
  Clinton      1      0        1
  Trump        0      0        0
  Not vote     0      0        2

, , Trump = Approve, Gender = Male, Age = 35 to 44 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        2      5        0
  Not vote     0      0        0
```

```
, , Trump = Disapprove, Gender = Male, Age = 35 to 44 years

          E2012
E2016      Obama Romney Not vote
  Clinton     12      0        0
  Trump        1      0        1
  Not vote     1      0        2

, , Trump = No Opinion, Gender = Male, Age = 35 to 44 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      1        0
  Not vote     0      0        0

, , Trump = Approve, Gender = Female, Age = 35 to 44 years

          E2012
E2016      Obama Romney Not vote
  Clinton      3      0        1
  Trump        3      2        0
  Not vote     0      0        0

, , Trump = Disapprove, Gender = Female, Age = 35 to 44 years

          E2012
E2016      Obama Romney Not vote
  Clinton      9      0        0
  Trump        0      0        0
  Not vote     0      0        0

, , Trump = No Opinion, Gender = Female, Age = 35 to 44 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      1        0
  Not vote     0      1        3

, , Trump = Approve, Gender = Male, Age = 45 to 54 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      9        0
  Not vote     0      0        0

, , Trump = Disapprove, Gender = Male, Age = 45 to 54 years

          E2012
```

```
E2016      Obama Romney Not vote
  Clinton      8      0        0
  Trump        1      1        0
  Not vote     1      0        1

, , Trump = No Opinion, Gender = Male, Age = 45 to 54 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      1        0
  Not vote     0      0        2

, , Trump = Approve, Gender = Female, Age = 45 to 54 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      5        1
  Not vote     1      0        0

, , Trump = Disapprove, Gender = Female, Age = 45 to 54 years

          E2012
E2016      Obama Romney Not vote
  Clinton     11      0        0
  Trump        0      0        1
  Not vote     0      0        1

, , Trump = No Opinion, Gender = Female, Age = 45 to 54 years

          E2012
E2016      Obama Romney Not vote
  Clinton      2      0        0
  Trump        0      0        0
  Not vote     0      0        1

, , Trump = Approve, Gender = Male, Age = 55 to 64 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        1      5        0
  Not vote     0      0        1

, , Trump = Disapprove, Gender = Male, Age = 55 to 64 years

          E2012
E2016      Obama Romney Not vote
  Clinton     13      0        0
  Trump        0      3        0
  Not vote     0      0        0
```

```
, , Trump = No Opinion, Gender = Male, Age = 55 to 64 years

          E2012
E2016      Obama Romney Not vote
  Clinton      1      0        0
  Trump        0      0        0
  Not vote     0      0        0

, , Trump = Approve, Gender = Female, Age = 55 to 64 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      3        1
  Not vote     0      0        0

, , Trump = Disapprove, Gender = Female, Age = 55 to 64 years

          E2012
E2016      Obama Romney Not vote
  Clinton     12      2        0
  Trump        0      0        0
  Not vote     0      0        0

, , Trump = No Opinion, Gender = Female, Age = 55 to 64 years

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      2        0
  Not vote     0      0        1

, , Trump = Approve, Gender = Male, Age = 65 years or over

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0     12        0
  Not vote     0      0        1

, , Trump = Disapprove, Gender = Male, Age = 65 years or over

          E2012
E2016      Obama Romney Not vote
  Clinton     11      2        0
  Trump        0      1        0
  Not vote     0      0        0

, , Trump = No Opinion, Gender = Male, Age = 65 years or over

          E2012
```

```
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      2        0
  Not vote     0      0        1

, , Trump = Approve, Gender = Female, Age = 65 years or over

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        0      5        1
  Not vote     0      0        0

, , Trump = Disapprove, Gender = Female, Age = 65 years or over

          E2012
E2016      Obama Romney Not vote
  Clinton      7      0        0
  Trump        0      0        0
  Not vote     0      0        0

, , Trump = No Opinion, Gender = Female, Age = 65 years or over

          E2012
E2016      Obama Romney Not vote
  Clinton      0      0        0
  Trump        1      0        0
  Not vote     0      0        0
```

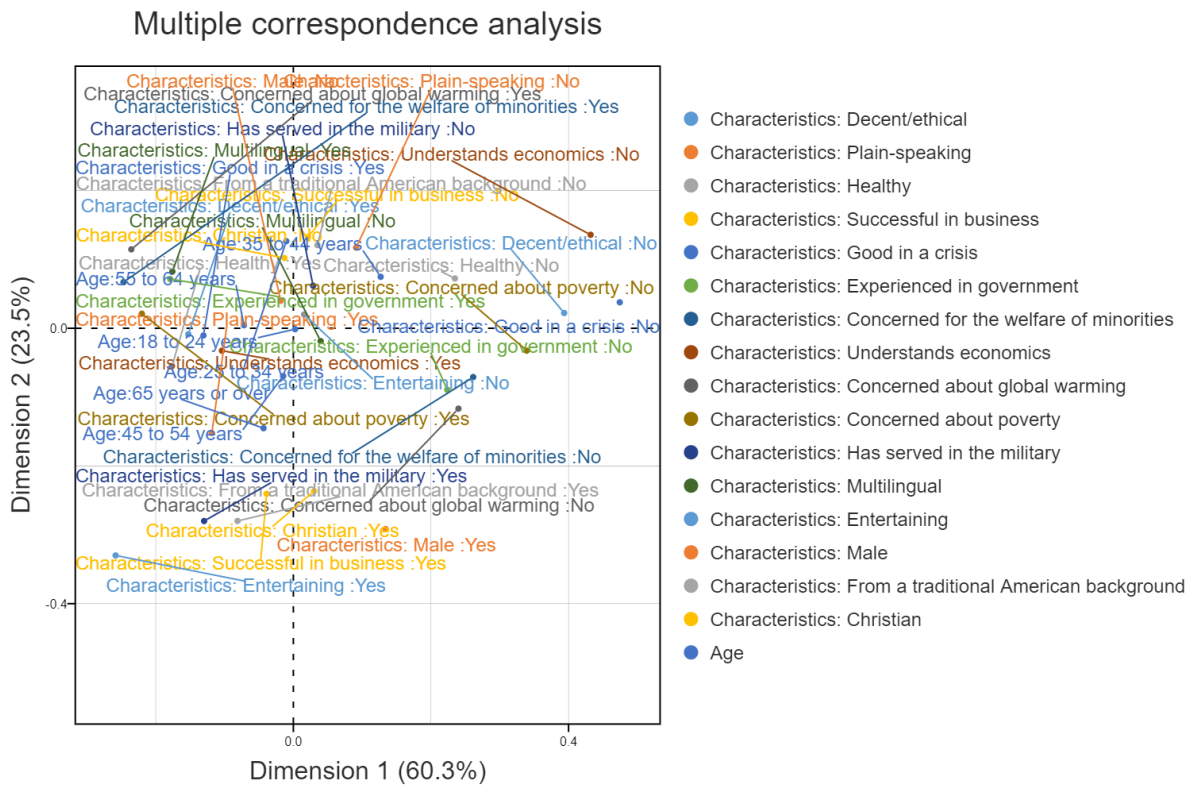
Inability to confidently evaluate associations

The next limitation relates to the interpretation of the relationships between the variables. As discussed in **HOW CORREPONDENCE ANALYSIS WORKS**, we can understand the association between labels from different variables by drawing a line from each label to the origin, and taking into account the lengths of these lines and the angle where they intersect with the origin. Unfortunately, with multiple correspondence analysis, there is no normalization that permits all such comparisons (see **Normalization and scaling**). Consequently, we always need to check the raw data. But, as just discussed, that is not so easy. Hence, with multiple correspondence analysis we have an increased risk of misinterpretation.

Further complicating the problem with looking at associations is that multiple correspondence analysis tends not to explain all the variance. In the map shown above, 16% of the variance is not explained. As we cannot inspect the data, this is a problem.

Messiness

The next problem relates to messiness. With more than five or six variables, the resulting maps are hard to use. As an example, look at the one below. What makes it so hard is that it plots every level of every categorical variable. Often this means that redundant information is plotted (e.g., the “yes” of a two-category, and, at the opposite side of the map, the “no” for the same variable). As shown below, a better approach is to use traditional correspondence analysis.



Unfocused

Last, multiple correspondence analysis produces unfocused analyses. In the analysis above, we are looking at the relationship between 17 variables relating to traits that people want in a US president, and age. The analysis treats all of the variables as being equally important. It will show the strongest relationships. That means that we will end up with a plot that explains how preference for the traits relates to age if and only if there is a very strong relationship between preferences for these characteristics and age.

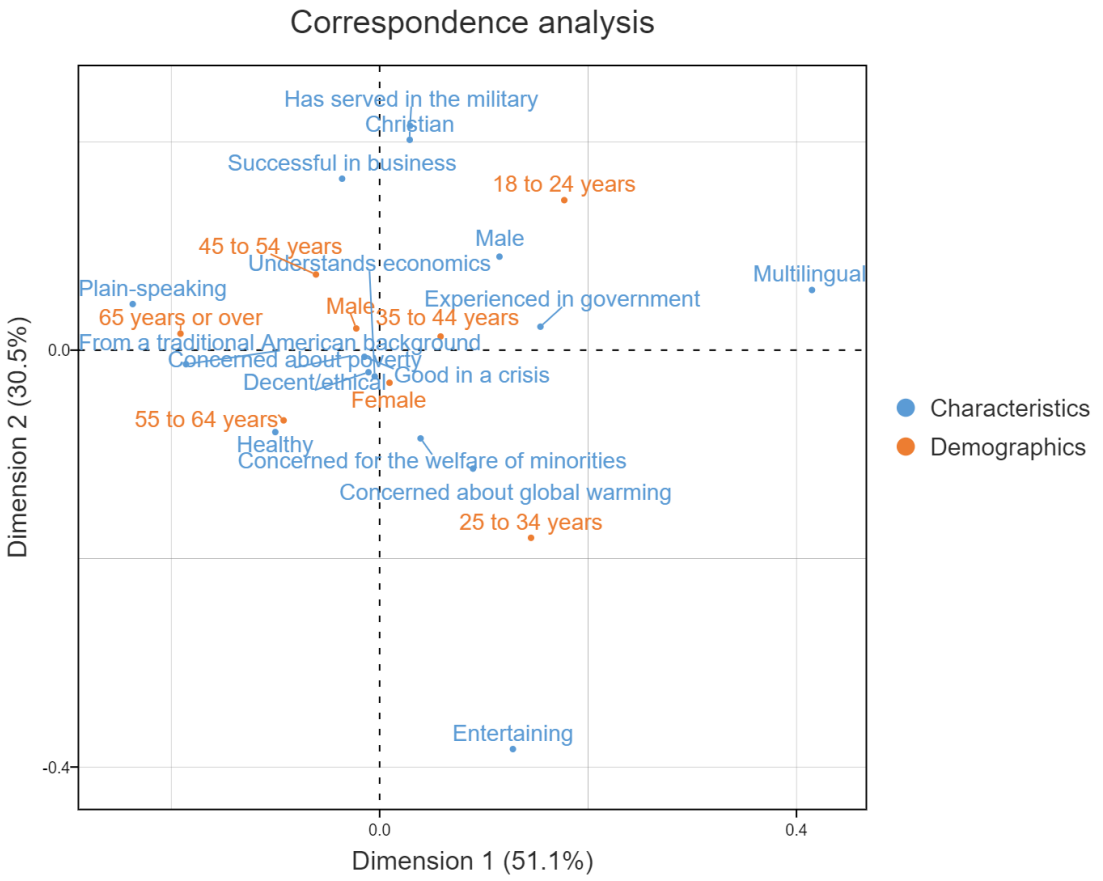
Correspondence analysis with multiple variables

The messy plot above represents a multi-dimensional table with 17 different dimensions! Sixteen of the dimensions have two levels each (i.e., whether a person mentioned a trait or not). The final dimension, age, has six levels. Thus, the input table has 2^16* 6 = 393,216 cells! Far too many to read. The table below uses the same 17 variables, but only has 96 cells. This is because rather than each of the trait variables being an extra dimension with two levels, we are just showing one of the levels and "stacking" them on top of each other as rows in the table. For example, 69% of people aged 18 to 24 said they wished for an American President that was Decent/Ethical, and 33% of people in this age band want the President to be Plain-speaking.

The multiple correspondence analysis shown in the previous section was based on the same data. However, it was based on the underlying variables. With traditional correspondence analysis, we first need to create a table. This is really what makes correspondence analysis so useful. We get to create a table in such a way that we focus on what we want to know. As this table compares age by the 16 variables, it will produce a plot that highlights the key relationships between age and the importance of these characteristics.

Column %	18 to 24 years	25 to 34 years	35 to 44 years	45 to 54 years	55 to 64 years	65 years or over
Decent/ethical	69%	72%	60% ▼	80%	75%	75%
Plain-speaking	33%	26% ▼	29% ▼	53%	54%	63% ▲
Healthy	44%	60%	41% ▼	58%	62%	75% ▲
Successful in business	44%	21% ▼	34%	42%	33%	42%
Good in a crisis	81%	75%	64% ▼	83%	85%	85%
Experienced in government	75% ▲	61%	55%	52%	48%	48%
Concerned for the welfare of minorities	53%	58%	43%	45%	63%	48%
Understands economics	75%	79%	76%	87%	85%	81%
Concerned about global warming	47%	61%	53%	43%	54%	42%
Concerned about poverty	61%	54%	55%	60%	67%	67%
Has served in the military	28%	11%	16%	22%	17%	19%
Multilingual	33% ▲	26%	19%	18%	12%	6% ▼
Entertaining	6%	12% ▲	2%	2%	6%	8%
Male	17%	14%	9%	15%	6%	13%
From a traditional American background	19%	23%	29%	32%	31%	44% ▲
Christian	47% ▲	19% ▼	26%	32%	27%	37%

The resulting plot (shown below) is a lot less messy. It tells us some things that are surprising on face-value. For example, the 25 to 34 years group are not so interested in a Christian president. As the analysis is based on a table, we can confirm this conclusion.



Although multiple correspondence analysis sounds better than correspondence analysis, the truth is the other way around. Multiple correspondence analysis is a technique that can be useful in special circumstances. By contrast, traditional correspondence analysis is applicable to the analysis of many different types of tables. As most data appears in a table at one time or another, correspondence analysis is a technique that can be widely applied.

Software

In Q and Displayr the process for creating multiple correspondence analysis is:

- Either:
    - In Q, select **Create > Dimension Reduction > Multiple Correspondence Analysis**.
    - In Displayr, select **Insert > Dimension Reduction > Multiple Correspondence Analysis**.
- Select the variables to be analyzed in the **Input variables** field.
- Check **Automatic**.

Correspondence analysis of multiple tables

> Traditional correspondence analysis can be used to analyze multiple tables at the same time. This is useful for comparing different sub-groups and for understanding trends.

This chapter shows how to use correspondence analysis to compare sub-groups. It focuses on one of the most interesting types of sub-groups: data at different points in time.[4] This is variously known as trend, tracking, longitudinal and time series data. The end-goal is a visualization showing key comparisons. Here we see how the positioning of different tech brands has changed over time. Thus, we are trying to do to two things at once:

- Create a correspondence analysis plot that shows key relationships (e.g., between brands and attributes, as above).
- Show how differences in these relationships vary by sub-group or over time.

The way that correspondence analysis proceeds with multiple tables is that we create tables of identical structure, such as the two below, which come from a technology study run in 2012 and repeated in 2017.

%	Fun	Innovative	Good customer service	Stylish	Easy-to-use	High quality	High performance	NET
Apple	46	66	26	59	38	53	50	79
Microsoft	17	39	18	10	34	30	34	65
Google	54	55	17	17	60	27	31	84
Sony	29	39	19	40	36	57	46	78
Yahoo	28	17	7	6	29	10	10	53
Samsung	18	28	17	29	30	37	30	68
LG	18	26	14	28	28	29	23	69
NET	100	100	100	100	100	100	100	100

Technology brand associations - 2012

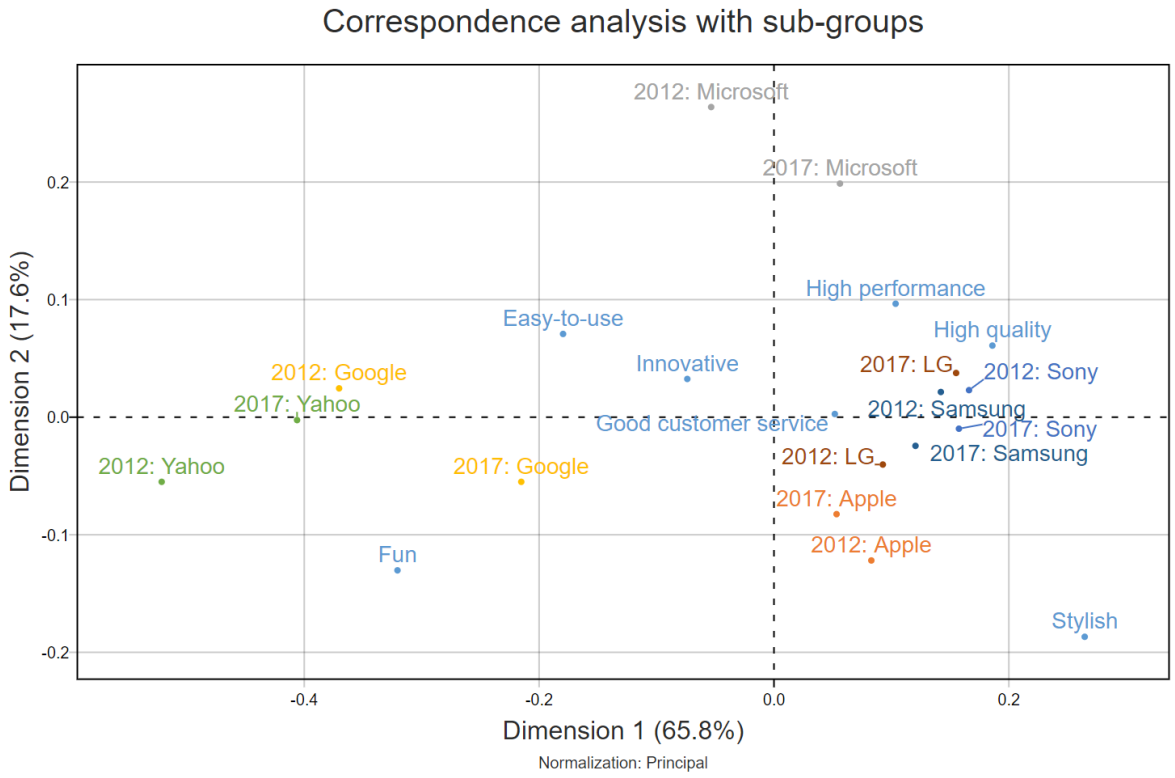
%	Fun	Innovative	Good customer service	Stylish	Easy-to-use	High quality	High performance	NET
Apple	64	75	51	69	59	72	66	86
Microsoft	22	43	21	20	38	46	45	71
Google	63	59	27	32	58	40	42	86
Sony	25	28	18	36	34	48	36	68
Yahoo	14	9	6	3	14	7	7	39
Samsung	29	50	30	52	51	49	46	77
LG	16	28	18	31	35	38	29	72
NET	100	100	100	100	100	100	100	100

Technology brand associations - 2017

[4] That is, one sub-group is the respondents that provided the data at one point at time, and another sub-group is those providing data at the other point in time.

Then, we splice them together to form a single table, joining them in such a way as to ensure that the resulting analysis focuses on what is of interest. The table below, when analyzed using correspondence analysis, will plot a position for 2012 and another for 2017, for each brand.

	Fun	Innova-tive	Good customer service	Stylish	Easy-to-use	High qual-ity	High perfor-mance
2012: Apple	46%	66%	26%	59%	38%	53%	50%
2012: Microsoft	17%	39%	18%	10%	34%	30%	34%
2012: Google	54%	55%	17%	17%	60%	27%	31%
2012: Sony	29%	39%	19%	40%	36%	57%	46%
2012: Yahoo	28%	17%	7%	6%	29%	10%	10%
2012: Samsung	18%	28%	17%	29%	30%	37%	30%
2012: LG	18%	26%	14%	28%	28%	29%	23%
2017: Apple	64%	75%	51%	69%	59%	72%	66%
2017: Microsoft	22%	43%	21%	20%	38%	46%	45%
2017: Google	63%	59%	27%	32%	58%	40%	42%
2017: Sony	25%	28%	18%	36%	34%	48%	36%
2017: Yahoo	14%	9%	6%	3%	14%	7%	7%
2017: Samsung	29%	50%	30%	52%	51%	49%	46%
2017: LG	16%	28%	18%	31%	35%	38%	29%



Traditional correspondence analysis is then used to analyze the resulting table. The output is shown above. The further apart two time periods are for a brand, the greater the difference between the sub-groups should be.

As always, it pays to confirm any key conclusions by checking the actual data. This is particularly important when comparing sub-groups, as sometimes the explanations are a bit more nuanced than is obvious.

Looking at the brands on left, we can see that both Google and Yahoo have moved towards the right (i.e., to quality, high performance, and style) in the past five years. However, the reasons are completely different. Google has improved. Yahoo has gone backwards in everything, but as its performance and style were already close to rock-bottom, they could not move much further and thus, in relative terms, Yahoo improved!

Two additional enhancements are:

- Only showing statistically significant differences. This is discussed at the end of the blog post Using Correspondence Analysis to Compare Sub-Groups and Understand Trends.
- Using arrow to show trends (see **Trend**).

Software

In Q and Displayr there is no need to create a single combined table, as the software does this automatically:

- Create the tables that you wish to compare. These need to have the same dimension (or, if they have different dimension, manually merge them and then use the standard correspondence analysis approach described in the earlier chapter).
- Either:
    - In Q, select **Create > Dimension Reduction > Correspondence Analysis of a Table**.
    - In Displayr, select **Insert > Dimension Reduction > Correspondence Analysis of a Table**.
- Select the *tables* in the **Input table(s)** field.
- Check **Automatic**.

When not to apply correspondence analysis

Correspondence analysis is useful when you have a table with the following characteristics:

- at least two rows of data
- at least two columns of data
- no missing data
- no negative values
- all the data has the same scale.

The only hard bit of this to understand is “same scale”, which is the focus of the examples in this chapter.

Tables containing multiple statistics

The table below shows both counts and column percentages. The data here is clearly on two different scales, making correspondence analysis inappropriate. We could scale the counts by turning them into percentages, but then we would just have the same data twice, which would be pointless.

Column % Count	18 to 24	25 to 39	40 to 49	50 to 64	65 or more	NET
Coca-Cola	64% 63	49% 131	28% 44	38% 84	34% 19	43% 341
Diet Coke	2% 2	12% 33	15% 24	8% 18	21% 12	11% 89
Coke Zero	10% 10	20% 53	21% 32	18% 40	14% 8	18% 143
Pepsi	6% 6	7% 18	21% 32	6% 13	5% 3	9% 72
Diet Pepsi	0% 0	1% 2	2% 3	5% 10	9% 5	3% 20
Pepsi Max	15% 15	12% 31	10% 15	22% 49	16% 9	15% 119
Dislike all cola	3% 3	0% 0	0% 0	1% 3	0% 0	1% 6
Don't care	0% 0	0% 0	4% 6	2% 4	0% 0	1% 10
NET	100% 99	100% 268	100% 156	100% 221	100% 56	100% 800

Multiple non-similar tables spliced together (bad, unless scaled)

The table below, which shows counts of cola preference by age and gender, would not be great for correspondence analysis. Why? The problem is that the data are not all on the same scale. There is an easy way to see this. If the data is all the same scale, it means that it is meaningful to sort the table by any of its rows and columns. If we were to sort this table by the first row, we would get Male and Female appearing first, because they have larger base sizes, and not because the sorting would be meaningful.

	18 to 24	25 to 39	40 to 49	50+	Male	Female	NET
Coca-Cola	63	131	44	103	169	172	341
Diet Coke	2	33	24	30	34	55	89
Coke Zero	10	53	32	48	59	84	143
Pepsi	6	18	32	16	45	27	72
Diet Pepsi	0	2	3	15	7	13	20
Pepsi Max	15	31	15	58	71	48	119
Dislike all cola	3	0	0	3	3	3	6
Don't care	0	0	6	4	7	3	10
NET	99	268	156	277	395	405	800

This gives us some insight into how to fix the problem. We need to transform the data in some way so that it is all on the same scale. We can achieve this by dividing each number by the column total (the NET at the bottom of the table), which gives us the table below. With this table, it makes sense to sort by the first row (which reveals that Coca-Cola preferences differ much more widely by age than by gender). It would also be appropriate to sort the table by any of the columns. Thus, it can be analyzed by traditional correspondence analysis.

	18 to 24	25 to 39	40 to 49	50+	Male	Female	NET
Coca-Cola	64%	49%	28%	37%	43%	42%	43%
Diet Coke	2%	12%	15%	11%	9%	14%	11%
Coke Zero	10%	20%	21%	17%	15%	21%	18%
Pepsi	6%	7%	21%	6%	11%	7%	9%
Diet Pepsi	0%	1%	2%	5%	2%	3%	3%
Pepsi Max	15%	12%	10%	21%	18%	12%	15%
Dislike all cola	3%	0%	0%	1%	1%	1%	1%
Don't care	0%	0%	4%	1%	2%	1%	1%
NET	100%	100%	100%	100%	100%	100%	100%

The next example has also been constructed using two tables. The last column shows the average attitude score for each brand, collected on a 5-point scale. Hopefully it is easy to see that the data in this table is not on the same scale, making it inappropriate for correspondence analysis.

	18 to 24	25 to 39	40 to 49	50 to 64	65 or more	Attitude
Coca-Cola	66	49	29	39	34	4
Diet Coke	2	12	16	8	21	3
Coke Zero	10	20	21	19	14	3
Pepsi	6	7	21	6	5	3
Diet Pepsi	0	1	2	5	9	3
Pepsi Max	16	12	10	23	16	3
NET	100	100	100	100	100	19

The next table shows the same data again, but “fixed”, so that it adds up to 100. Is this table OK? No. The best way to appreciate the problem is to focus on Diet Pepsi. Diet Pepsi has the lowest score of any of the brands on Attitude. However, if we read across the Diet Pepsi row, we see that Diet Pepsi gets its highest score for Attitude, so if we did run a correspondence analysis we would conclude (incorrectly) that Diet Pepsi was really strong on Attitude, even though this is not the case, as in relative terms Diet Pepsi’s attitude score is strong compared to the other columns in the table.

	18 to 24	25 to 39	40 to 49	50+	NET	Attitude
Coca-Cola	66	49	29	38	43	21
Diet Coke	2	12	16	11	11	15
Coke Zero	10	20	21	18	18	17
Pepsi	6	7	21	6	9	17
Diet Pepsi	0	1	2	6	3	14
Pepsi Max	16	12	10	21	15	17
NET	100	100	100	100	100	100

Table of correlations

The next table shows correlations between two sets of variables. Correspondence analysis will not work here as we have negative values.

Correlation	Coca-Cola	Diet Coke	Coke Zero	Pepsi	Diet Pepsi	Pepsi Max	SUM
Coca-Cola - When 'out and about'	.48	-.23	-.13	-.01	-.25	-.23	-.13
Diet Coke - When 'out and about'	-.19	.51	.11	-.11	.24	.06	.21
Coke Zero - When 'out and about'	-.10	.17	.41	-.10	.04	.09	.18
Pepsi - When 'out and about'	.10	.00	-.03	.32	.04	.07	.15
Diet Pepsi - When 'out and about'	-.18	.17	.01	.02	.44	.09	.16
Pepsi Max - When 'out and about'	-.12	.04	.07	.16	.10	.52	.25
Coca-Cola - When 'at home'	.51	-.30	-.16	-.03	-.27	-.29	-.19
Diet Coke - When 'at home'	-.22	.55	.08	-.13	.24	.04	.19
Coke Zero - When 'at home'	-.12	.24	.60	-.15	.07	.13	.27
Pepsi - When 'at home'	.10	-.03	-.04	.34	.03	.09	.14
Diet Pepsi - When 'at home'	-.16	.18	-.01	.02	.41	.07	.16
Pepsi Max - When 'at home'	-.12	.07	.09	.17	.13	.54	.28
SUM	.03	.31	.36	-.03	.12	.19	.32

We can fix this by adding a 1 to every cell in the table, which would mean that there are no longer negative results, making it practical to apply correspondence analysis. Sure, there are perhaps better techniques, such as canonical correlation analysis, but we can extract insights from such a table using correspondence analysis.

INTERPRETATION

This section focuses on interpretation. The three chapters in this section describe:

- The correct interpretation of traditional correspondence analysis.
- The math of correspondence analysis.
- Normalization and scaling.

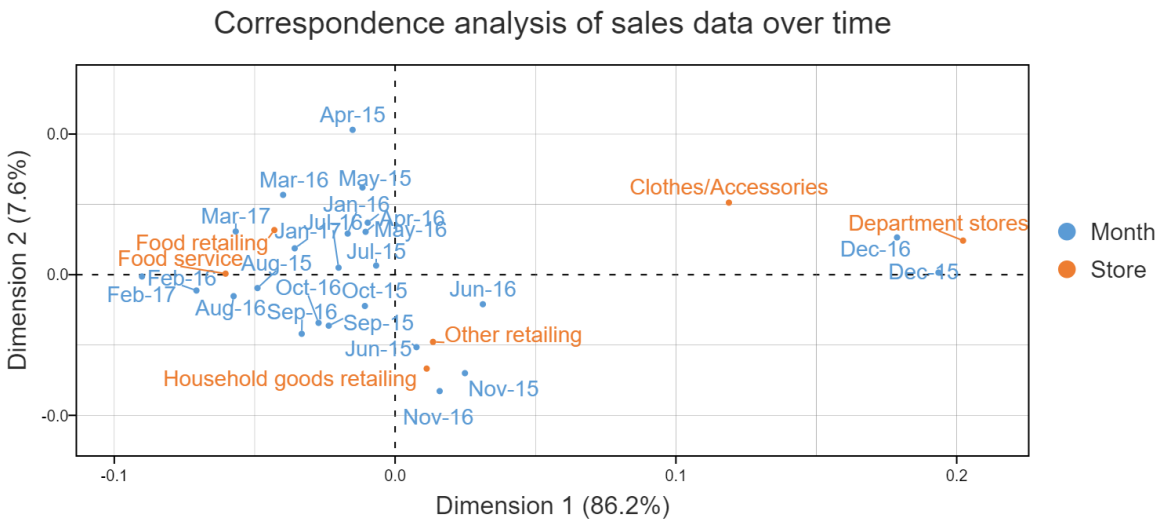
The correct interpretation of correspondence analysis maps

> Although correspondence analysis is a useful simplification of large tables, the correct interpretation is not obvious. This chapter works through a series of examples, illustrating key concepts in the interpretation of correspondence analysis.

1. Check conclusions using the input data

The key to correctly interpreting correspondence analysis is to check any important conclusions by referring back to the original data. In this chapter I list nine other things to think about when interpreting correspondence analysis. But, so long as you always remember this first rule, you will not go wrong.

The reason for this rule is illustrated in the example below. It shows 24 months of sales data by different retail categories. The visualization shows that Department stores are associated with December (i.e., Christmas, Dec-15 and Dec-16). We can see that Food retailing is on the opposite side of the map, which most people would interpret as meaning that Food retailing sales are lower in December.



Now, look at the actual data, shown below. Even though Food retailing is a long way from December on the map:

- Food retailing has the highest sales in December of any of the categories.
- Food retailing's biggest month is December.

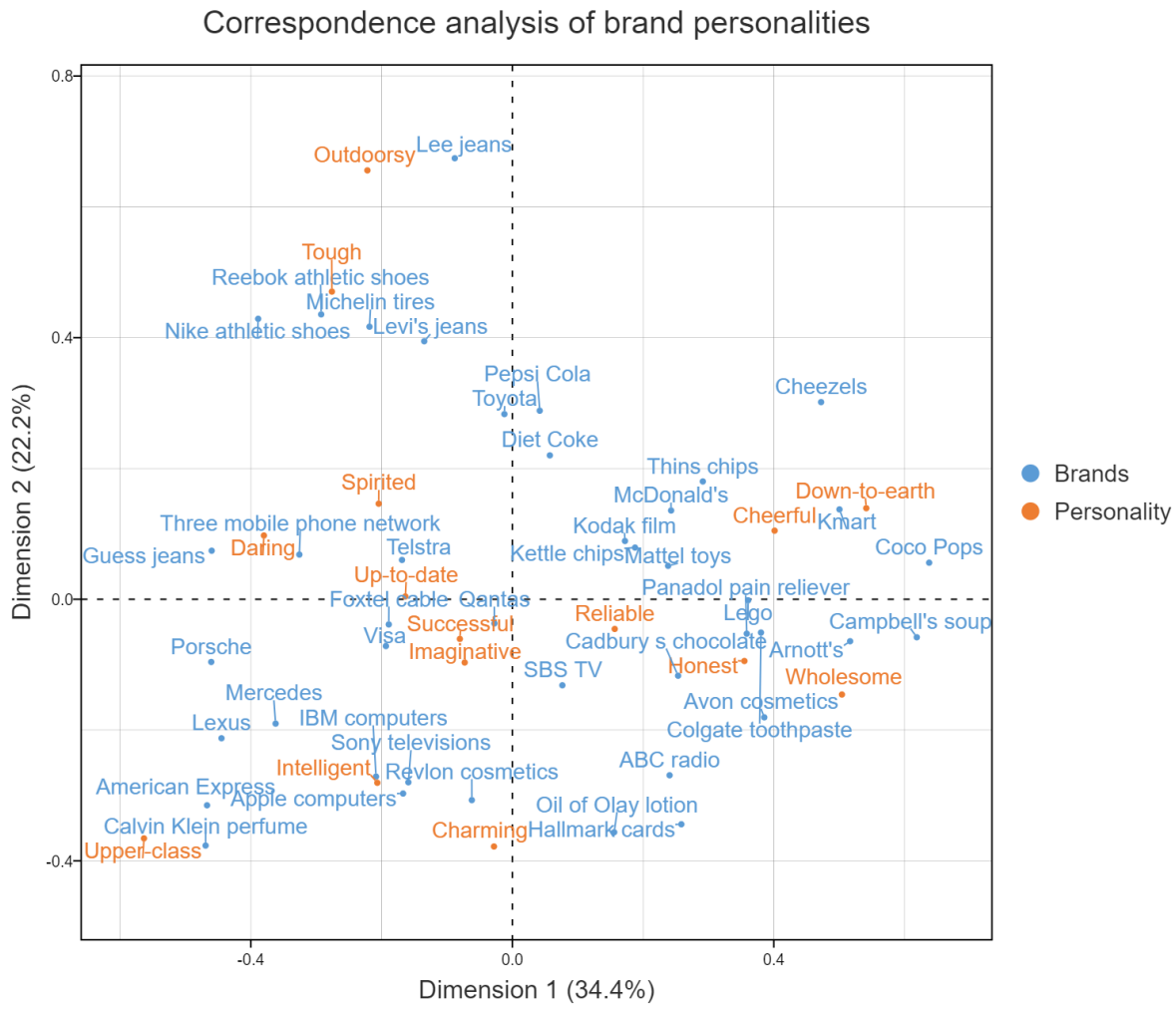
	Food retailing	Household goods retailing	Clothes/ Accessories	Department stores	Other retailing	Food service
Apr-15	9538	3714	1793	1442	3059	3267
May-15	9707	3977	1891	1429	3215	3314
Jun-15	9218	4319	1775	1481	3244	3258
Jul-15	9718	4186	1789	1541	3337	3445
Aug-15	9835	4160	1697	1332	3380	3421
Sep-15	9646	4274	1790	1400	3455	3444
Oct-15	10319	4568	1889	1566	3580	3526
Nov-15	10129	4693	1916	1731	3825	3491
Dec-15	11731	5736	3080	2914	4643	3820
Jan-16	10245	4377	1876	1519	3305	3432
Feb-16	9557	3980	1599	1156	3257	3187
Mar-16	10354	4097	1781	1452	3399	3435
Apr-16	9728	4065	1925	1451	3356	3452
May-16	9815	4093	1927	1450	3429	3431
Jun-16	9517	4357	1967	1596	3414	3314
Jul-16	9929	4225	1876	1468	3493	3573
Aug-16	10042	4239	1806	1294	3562	3648
Sep-16	10006	4469	1897	1394	3602	3696
Oct-16	10483	4697	1938	1497	3643	3717
Nov-16	10436	4874	2057	1684	4051	3679
Dec-16	12230	5782	3331	2850	4860	4047
Jan-17	10432	4464	1940	1429	3422	3621
Feb-17	9575	3898	1559	1092	3231	3261
Mar-17	10510	4197	1827	1370	3590	3619

How can this be? The data seems to say the exact opposite of visualization? If you have read **HOW CORREPONDENCE ANALYSIS WORKS**, you should understand that this is because correspondence analysis is all about the relativities. If we dig deeper into the data we can see that the map above does make sense, once you know how to read it.

While Food retailing does peak at Christmas, its sales are only 19% above its average monthly sales. By contrast, Department store sales spike to 85% above average in December. This is what correspondence analysis is trying to show us. Correspondence analysis does not show us which rows have the highest numbers, nor which columns have the highest numbers. It instead shows us the relativities. If your interest is instead on which categories sell the most, or how sales change over time, you are better off plotting the raw data than using correspondence analysis.

2. The further things are from the origin, the more discriminating they are

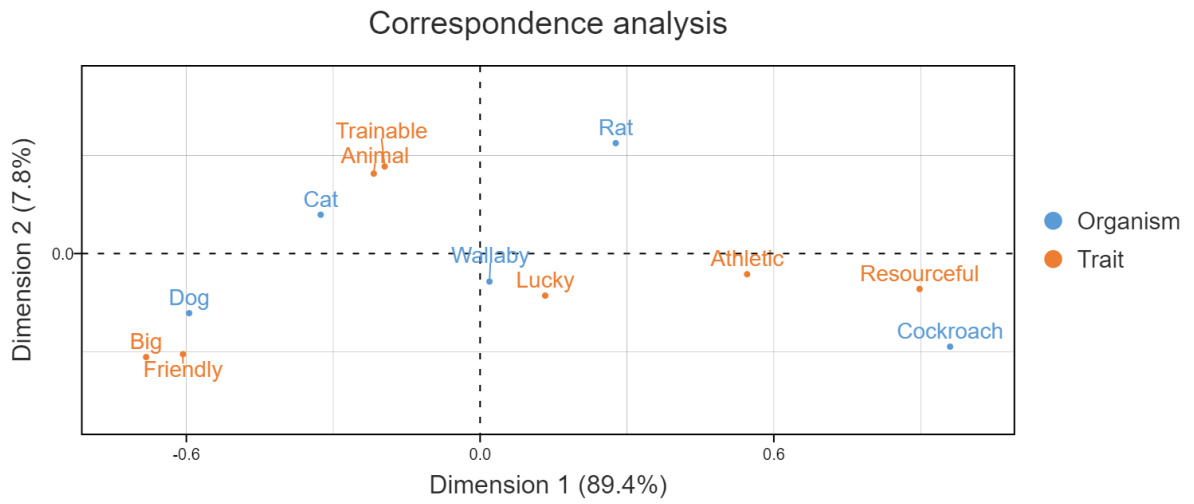
The correspondence analysis plot below is from a big table consisting of 42 rows, each representing a different brand, and 15 columns. You can see the original data here. Correspondence analysis has greatly simplified the story in the data. As you hopefully remember from school, the origin is where the x- and y-axes are both at 0. It is shown below as the intersection of two dashed lines. The further labels are from the origin, the more discriminating they are. Thus, Lee Jeans (at the top) is highly differentiated. Similarly, Outdoorsy is a highly discriminating attribute.



3. The closer things are to origin, the less distinct they *probably* are

In the map above, we see that Qantas is bang smack in the middle of the visualization. Thus, the conclusion probably is that it is not differentiated based on any of the data in the study. I explain the use of the weasel-word “probably” in the next section.

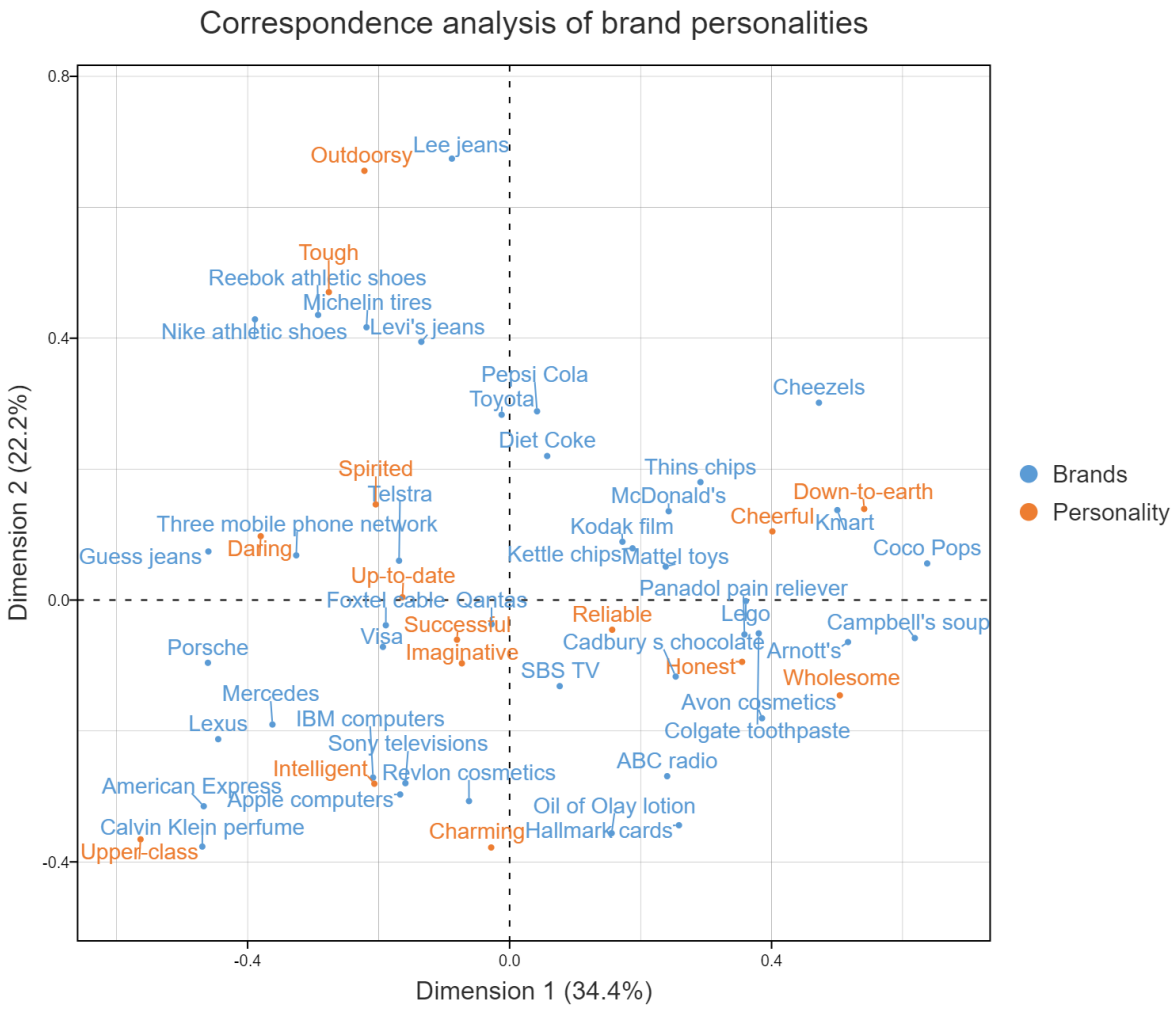
Here is another example. In the center of the map we have Wallaby and Lucky. Does this mean wallabies are lucky animals? No. They get hit by cars a lot. If you follow rugby, you will know that 99 times out of 100 a Wallaby is no match for even a Kiwi. If you look at the table below, you can see that the Wallaby is pretty average on all the variables being measured. As it has nothing that differentiates it, the result is that it is in the middle of the map (i.e., near the origin). Similarly, Lucky does not differentiate, so it is also near the center. That they are both in the center tells us that they are both indistinct, and that is all that they have in common (in the data).



	Big	Athletic	Friendly	Trainable	Resour-ceful	Animal	Lucky
Dog	80	20	90	90	5	100	40
Cat	50	40	40	70	10	100	40
Rat	10	70	20	90	80	99	40
Cockroach	0	80	2	20	95	20	40
Wallaby	35	52	38	47	48	80	40

4. The more variance explained, the fewer insights will be missed

The correspondence analysis of the brand personality data is reproduced below. You will hopefully recall that Qantas being in the middle meant that it was probably not differentiated based on the data. Why “probably”? If you sum up the proportion of variance explained by horizontal and vertical dimensions (shown in the axis labels), we see that visualization displays 57% of the variance in the data. And, remember, this is only 57% of the variance in the relativities. So, a lot of the data has been left out of the summary. Qantas may be highly differentiated on some dimension that is less relevant for most of the brands; the only way to know for sure is to check the data.

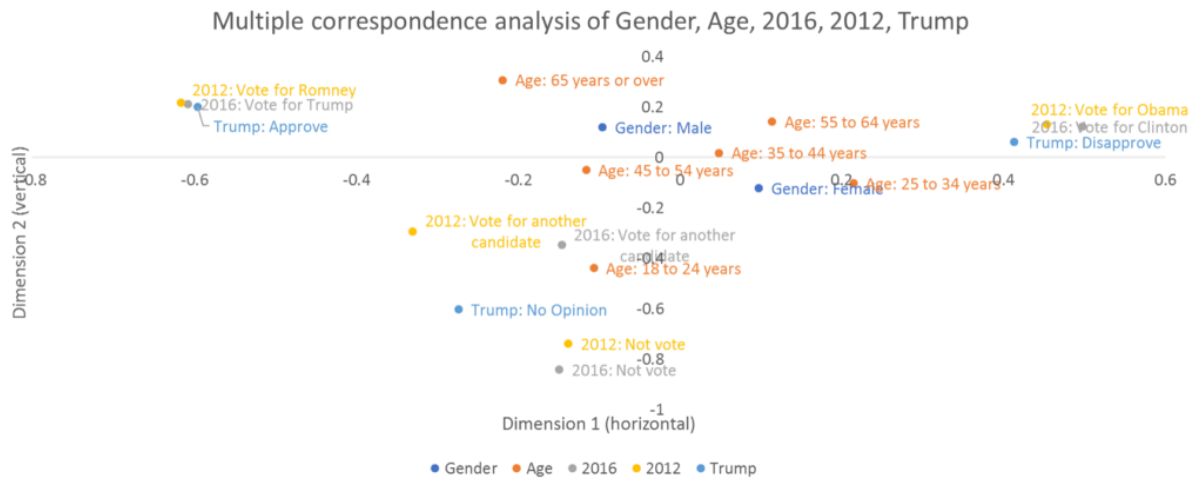


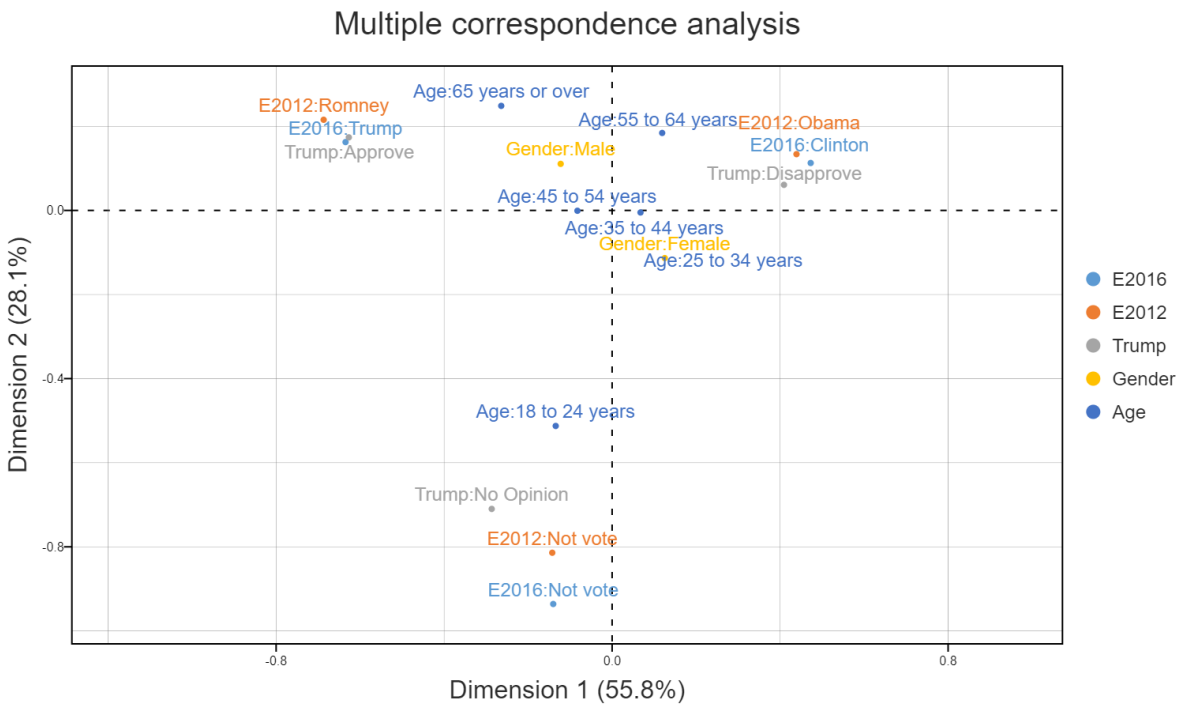
In fairness to correspondence analysis, it is a great achievement for the map to explain 57% of the variation with such a large input table. To represent all the relativities of this table requires 14 dimensions, but we have only plotted two. Correspondence analysis is not the problem. The problem is the quantity of the data. The more data, the greater the chance that any good summary will miss important details.

5. Proximity between row labels probably indicates similarity (if properly normalized)

As discussed in some detail in **HOW CORREPONDENCE ANALYSIS WORKS**, we should be able to gauge the similarity of row labels based on their distance on the map (i.e., their proximity). "Should" is another weasel word! Why? Three things are required in order for this to be true:

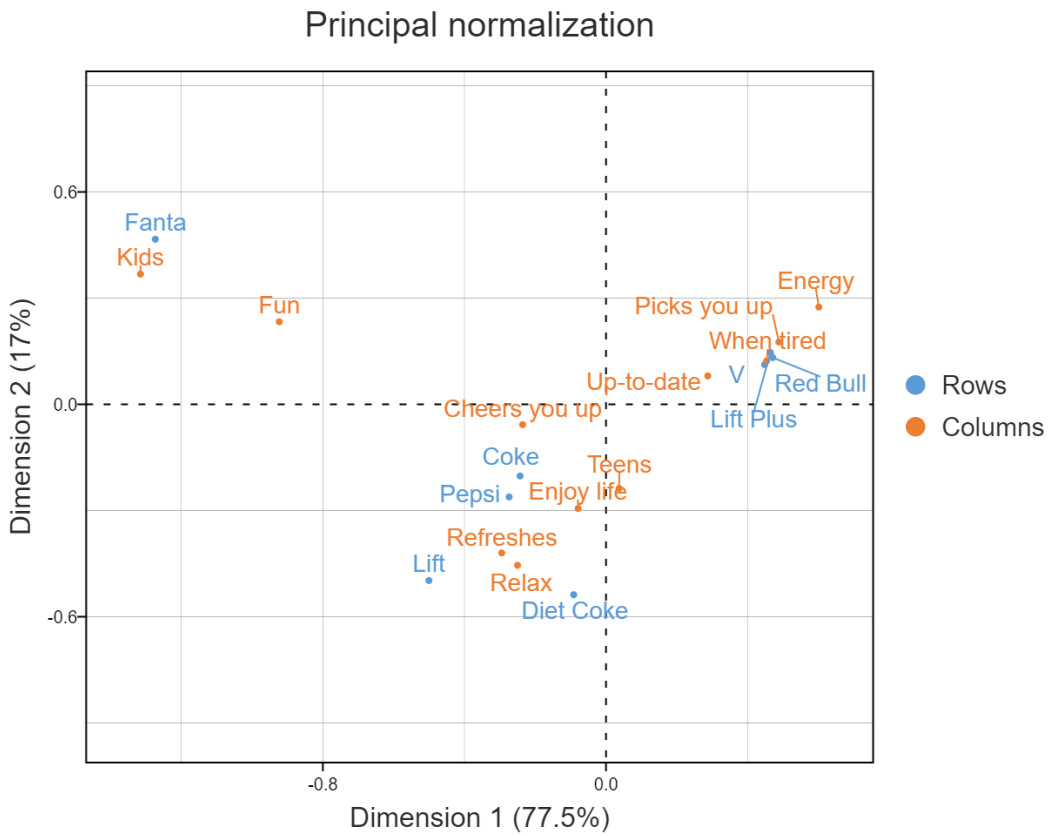
1. We need to be explaining a high proportion of variance in the data. If we are not, there is always the risk that the two row labels are highly distinct but are still shown on the map as if not distinct.
2. The *normalization*, which is a technical option in correspondence analysis software, needs to have been set to either *principal*, *row principal*, or *row principal (scaled)* (see **Normalization and scaling**).
3. The *aspect ratio* of the map needs to be fixed at 1. That is, the horizontal and vertical coordinates of the map need to be drawn to the same scale. If your maps are in Excel or, as in the example below, PowerPoint, you may well have a problem. In the chart below, the really big pattern is that there is an enormous gap between the pro-Trump camp, on the far left, and the pro-Clinton camp on the far right. If you have even a passing understanding of American politics, this will make sense. However, if you look at the scale of the labels on the x- and y- axes you will see a problem. A distance of 0.2 on the horizontal is equivalent to a distance of 0.6 on the vertical. The map below this has the aspect ratio set to 1, and it tells a different story. Yes, the pro- and anti-Trump camps are well apart, but the disenfranchised youth are now much more prominent.


Multiple correspondence analysis of Gender, Age, 2016, 2012, Trump
Age: 65 years or over
2012: Vote for Romney
2016: Vote for Trump
Trump: Approve
Gender: Male
Age: 55 to 64 years
2012: Vote for Obama
2016: Vote for Clinton
Trump: Disapprove
Age: 35 to 44 years
Age: 45 to 54 years
Gender: Female
Age: 25 to 34 years
2012: Vote for another candidate
2016: Vote for another candidate
Age: 18 to 24 years
Trump: No Opinion
2012: Not vote
2016: Not vote
Dimension 2 (vertical)
Dimension 1 (horizontal)
Gender
Age
2016
2012
Trump

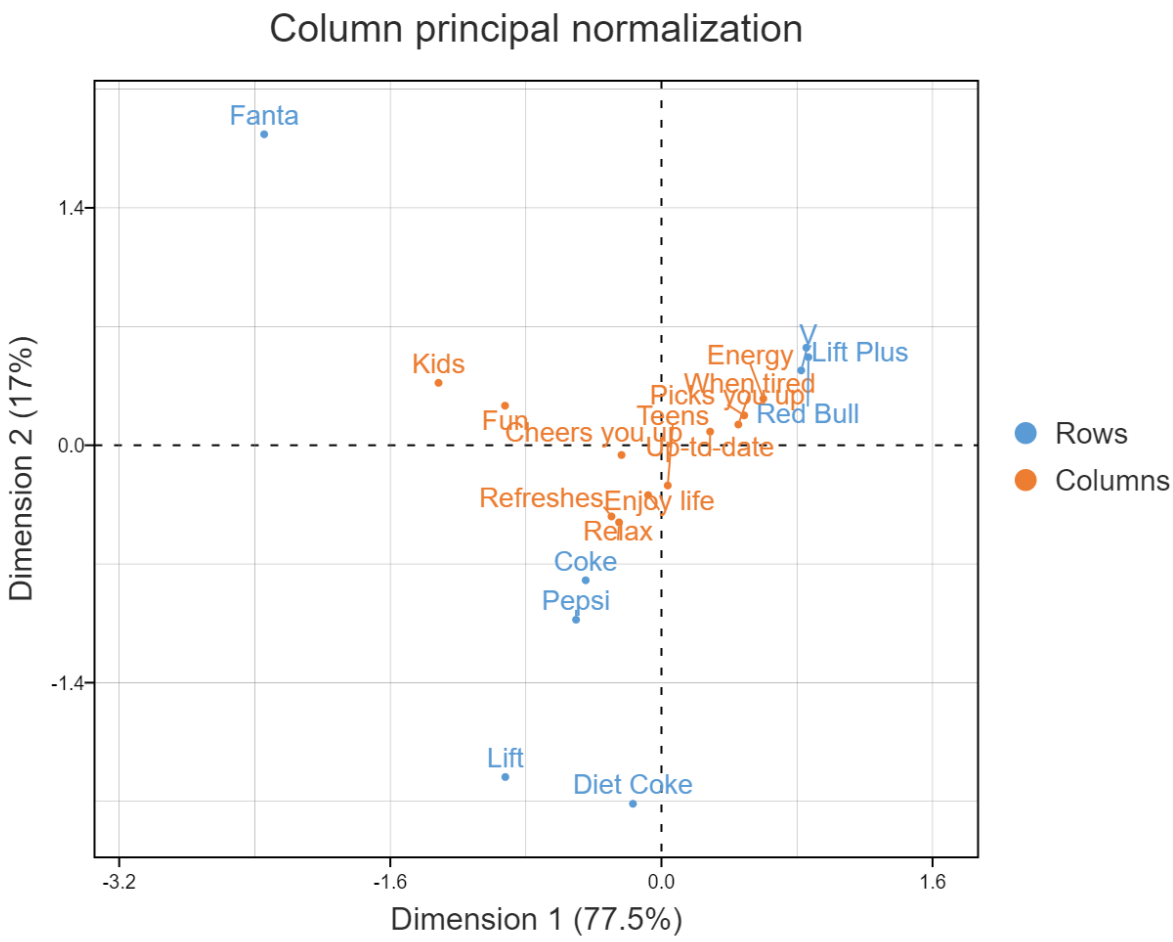

Multiple correspondence analysis
E2012:Romney
E2016:Trump
Trump:Approve
Age:65 years or over
Age:55 to 64 years
E2012:Obama
E2016:Clinton
Trump:Disapprove
Gender:Male
Age:45 to 54 years
Age:35 to 44 years
Gender:Female
Age:25 to 34 years
Age:18 to 24 years
Trump:No Opinion
E2012:Not vote
E2016:Not vote
Dimension 2 (28.1%)
Dimension 1 (55.8%)
E2016
E2012
Trump
Gender
Age

6. Proximity between column labels indicates similarity (if properly normalized)

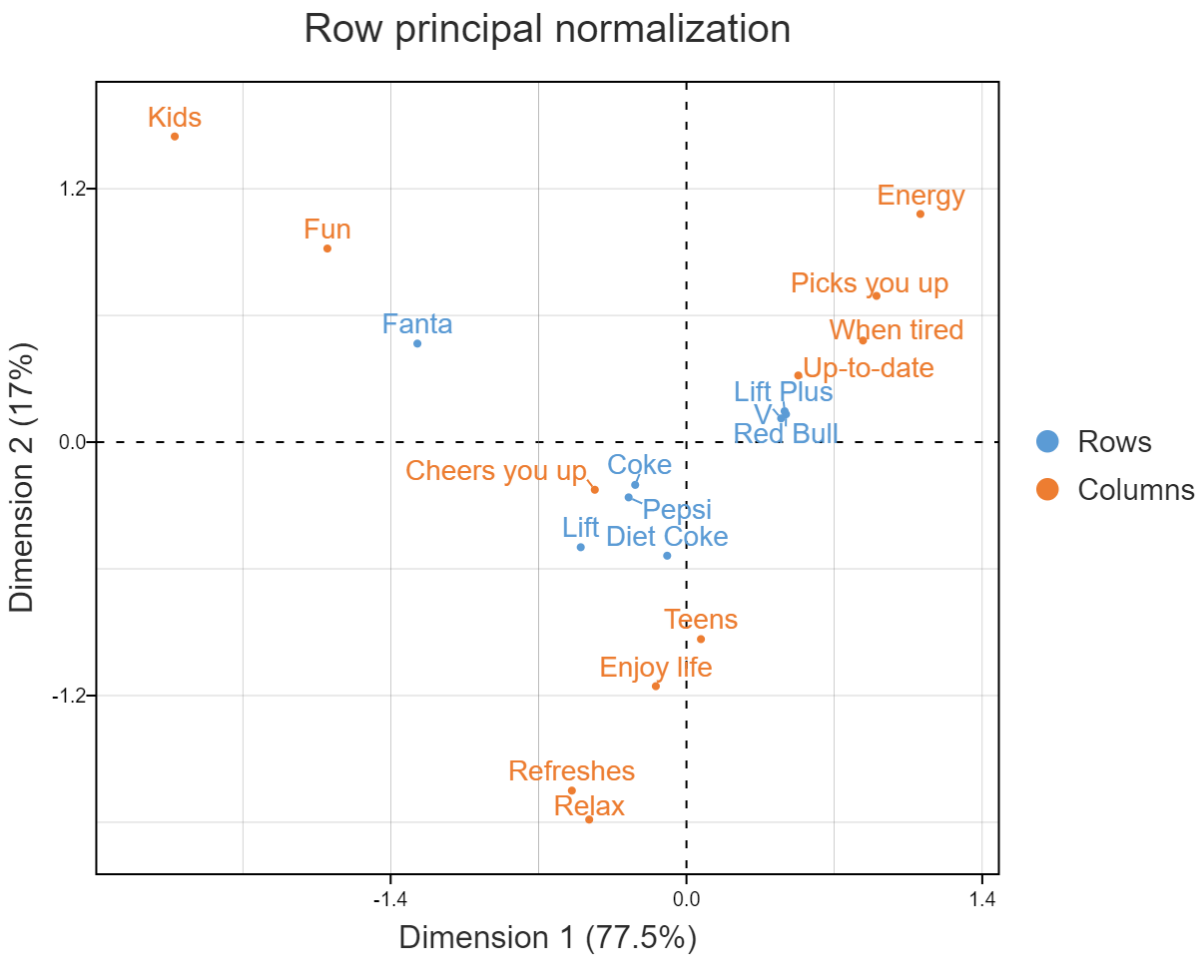
This is a repeat of the previous point, but applied to columns. Here, the normalization needs to be either *principal*, *column principal*, or *column principal (scaled)*. You may recall that to compare between rows, we need to be using either principal, row principal, or row principal (scaled) normalization. So, setting the normalization to principal seems like the obvious solution. But, before jumping to this conclusion, which has its own problems (as discussed in the next section), I will illustrate what these different normalization settings look like. The visualization below is based on the principal normalization. Principal is the default in some apps, such as Displayr and Q. However, it is not the default in SPSS, which means that comparing the distances between rows labels in a map created by SPSS with defaults is dangerous.



The plot below uses the *column principal normalization*. If you look very carefully, you will see that the positions of the column points are unchanged (although the map has been zoomed out). But, the positions of the row labels, representing the brands, have changed. There are two ways that the row labels positions have changed. First, they have been stretched out to be further from the origin. Second, the degree of stretching has been greater vertically. With the principal plot shown above, the horizontal differences for the row labels are, in relative terms, bigger. With the column principal shown below, the vertical differences are bigger. So, to repeat the point made a bit earlier: the distances between the column points are valid for both principal and column principal, but the distances between the row points are not correct in the column principal shown below.



The visualization below shows the *row principal normalization*. Now the distances between the row labels are meaningful and consistent with those shown in the principal normalization, but the differences between the column coordinates are now misleading.



7. If there is a small angle connecting a row and column label to the origin, they are probably associated

Take a look at the plot above. Would you say Lift is more strongly associated with Cheers you up or Relax? If you said Relax, you are interpreting the map correctly. As discussed in **HOW CORREPONDENCE ANALYSIS WORKS**, it is wrong to look at the distance between row labels and column labels. Instead, we should imagine a line connecting the row and column labels with the

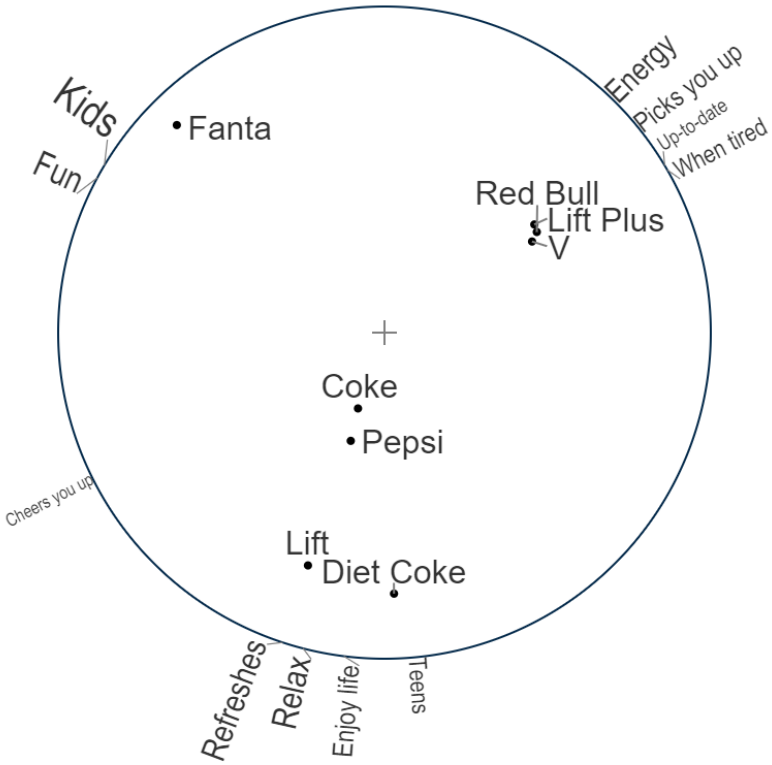
origin. The sharper the angle, the stronger the relationship. Thus, there is a strong relationship between Relax and Lift (although, if you look at the data shown below, you will see that Lift is very small, so it does not in any sense "own" Relax).

	Coke	V	Red Bull	Lift Plus	Diet Coke	Fanta	Lift	Pepsi
Kids	30%	2%	1%	1%	2%	45%	7%	8%
Teens	69%	46%	41%	24%	18%	13%	11%	22%
Enjoy life	50%	22%	19%	9%	7%	8%	6%	10%
Picks you up	29%	52%	45%	27%	3%	3%	2%	5%
Refreshes	28%	12%	7%	5%	4%	7%	12%	5%
Cheers you up	26%	12%	11%	6%	3%	11%	4%	4%
Energy	19%	55%	47%	28%	1%	2%	2%	3%
Up-to-date	28%	30%	29%	17%	3%	5%	2%	6%
Fun	35%	6%	5%	3%	2%	32%	3%	5%
When tired	25%	38%	33%	19%	3%	2%	1%	4%
Relax	21%	6%	4%	2%	2%	3%	4%	3%

If you have not yet had your coffee for the day, go get it now. We are at the hard bit. In the plot above, the angles are informative. However, interpreting the angles is only strictly valid when you have either row principal, column principal, row principal (scaled), column principal (scaled) or symmetrical (1/2) normalization. So, if we want to make inferences about the relationships between the rows and columns (e.g., brands and attributes in the examples above), we are better off not using the default principal normalization. This is really the yuckiest aspect of correspondence analysis. No one normalization is appropriate for everything. Or, stated from a glass half full perspective, our choice of normalization is really a choice of how we want to mislead the viewer!

This problem is made more complex by the tendency of people not to report the normalization. Fortunately, we can make an educated guess based on the dispersion of the points (if the rows points are all near the origin we probably have row principal, and vice versa for columns).

Depending on the day of the week I have two ways of dealing with this issue. Most of the time, my preference is to use the principal normalization, and remind viewers to check everything in the raw data. Sometimes though, where my interest is principally in the rows of a table, I use row principal and a *moonplot*. Distances between the brands are plotted inside of a circle and these distances are meaningful. The column labels are shown on the outside of the circle. They have the same angles as on the plot above. But now the font size represents what was previously indicated by the distance between the column labels and the origin. The beauty of this representation is that we can now compare distances between column labels and points, so the plot is much harder to misread, and we have no need to educate the reader about the whole reading of angles. The information regarding the relativities of the column labels is harder to gauge, but this is arguably beneficial, as the construction of the plot makes it clear that the focus is on the rows (brands).



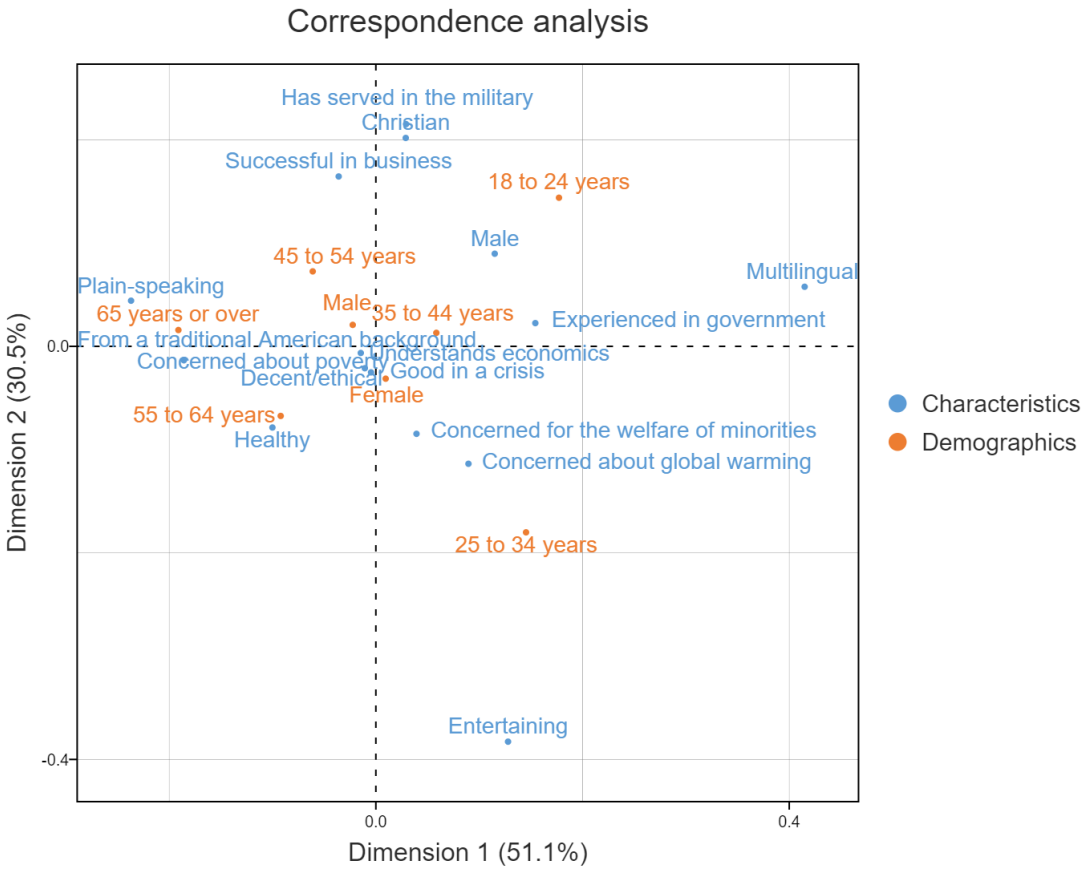
Moonplots are discussed again in the **VISUALIZATION** section of this eBook.

8. A row and column label are probably not associated if their angle to the origin is 90 degrees

In the moonplot above, if you draw a line connecting Red Bull to the origin, and back out to Kids, you will see that it is roughly a right-angle (90 degrees). This tells us that there is no association between Kids and Red Bull. Again, I have written “probably”. If you look at the data, shown in the table above, there is clearly a negative association. Remember, always look at the data!

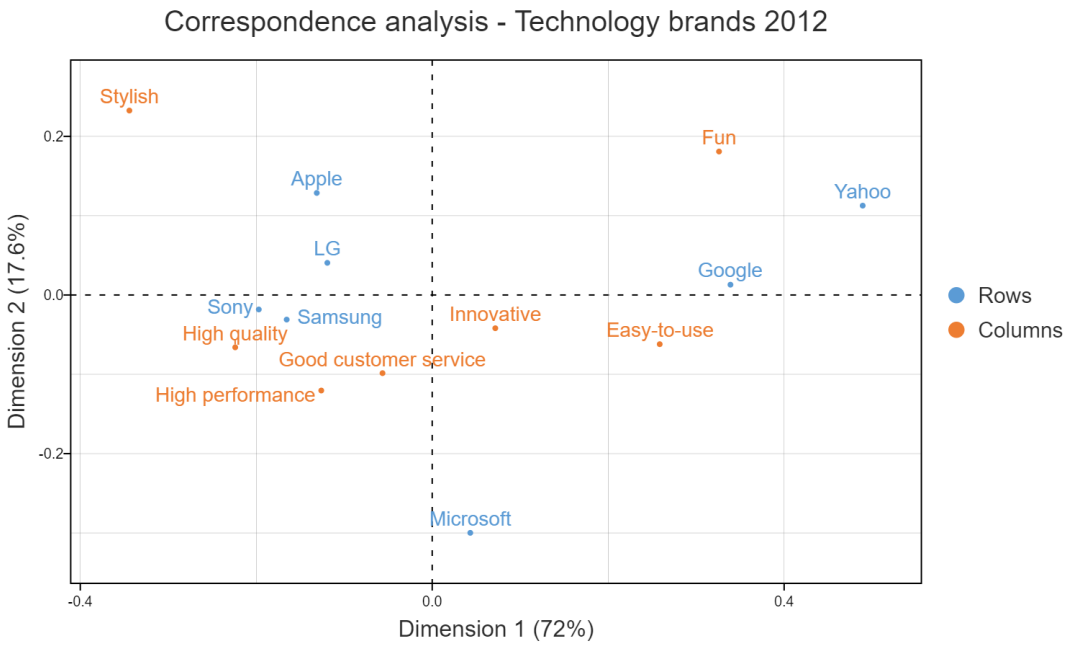
9. A row and column label are probably negatively associated if they are on opposite sides of the origin

The plot below shows the traits that people want in an American president by age. What do the 25 to 34 years group yearn for? The strongest association with Entertaining. What is the next strongest association? You may think it would be concern about global warming and minorities. This is not the case. The next strongest associations are negative ones: the 25 to 34 years are less keen on a Christian president, one who has been successful in business, and one who is plain-speaking. We can see this because these traits are on the opposite side of the origin, and are a long way from the origin, whereas the traits relating to global warming and welfare of minorities are all closer to the origin, and thus are less discriminating.



Column %	18 to 24 years	25 to 34 years	35 to 44 years	45 to 54 years	55 to 64 years	65 years or over	NET
Decent/ethical	69%	72%	60% ▼	80%	75%	75%	72%
Plain-speaking	33%	26% ▼	29% ▼	53%	54%	63% ▲	43%
Healthy	44%	60%	41% ▼	58%	62%	75% ▲	57%
Successful in business	44%	21% ▼	34%	42%	33%	42%	36%
Good in a crisis	81%	75%	64% ▼	83%	85%	85%	78%
Experienced in government	75% ▲	61%	55%	52%	48%	48%	56%
Concerned for the welfare of minorities	53%	58%	43%	45%	63%	48%	51%
Understands economics	75%	79%	76%	87%	85%	81%	81%
Concerned about global warming	47%	61%	53%	43%	54%	42%	50%
Concerned about poverty	61%	54%	55%	60%	67%	67%	61%
Has served in the military	28%	11%	16%	22%	17%	19%	18%
Multilingual	33% ▲	26%	19%	18%	12%	6% ▼	18%
Entertaining	6%	12% ▲	2%	2%	6%	8%	6%
Male	17%	14%	9%	15%	6%	13%	12%
From a traditional American background	19%	23%	29%	32%	31%	44% ▲	30%
Christian	47% ▲	19% ▼	26%	32%	27%	37%	30%
NET	100%	100%	100%	100%	100%	100%	100%

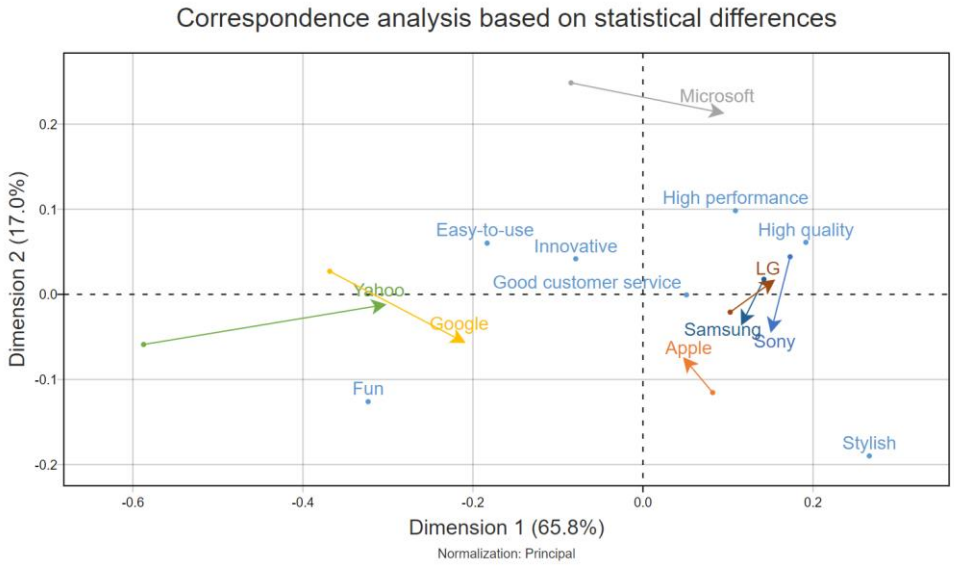
Here's another example. The correct way to read this visualization is that Yahoo is, in relative terms, stronger than Google on Fun. However, if you look at the raw data it shows that Google is much more fun than Yahoo (54% versus 28%). The reason that Yahoo has stronger association with Fun is that it is its second best performing attribute (with 29% for Easy-to-use). By contrast, while Google is twice as fun as Yahoo, it scores three times as high on High quality and High performance, which are on the opposite side of the map, and this is what drags Google away from Yahoo.



%	Fun	Innovative	Good customer service	Stylish	Easy-to-use	High quality	High performance	NET
Apple	46%	66% ↑	26%	59% ↑	38% ↓	53%	50%	79% ↓
Microsoft	17% ↓	39% ↑	18% ↑	10% ↓	34%	30%	34% ↑	65%
Google	54% ↑	55% ↑	17%	17% ↓	60% ↑	27% ↓	31% ↓	84% ↑
Sony	29% ↓	39% ↓	19%	40% ↑	36% ↓	57% ↑	46% ↑	78%
Yahoo	28% ↑	17%	7%	6% ↓	29% ↑	10% ↓	10% ↓	53%
Samsung	18% ↓	28% ↓	17%	29% ↑	30%	37% ↑	30%	68%
LG	18% ↓	26%	14%	28% ↑	28%	29%	23%	69%
NET	100%	100%	100%	100%	100%	100%	100%	100%

10. The further a point from the origin, the stronger their positive or negative association

The visualization below shows movement of Yahoo's perceptions from 2012 to 2017, with the arrow head showing 2017 and the base of the arrow showing 2012. The obvious way to read this is that Yahoo has become more fun, more innovative, and easier-to-use. However, such a conclusion is misplaced.



A better interpretation is:

- In 2012, the angle formed by connecting the base of Yahoo to the origin and back to Fun is very small, which tells us that they are associated.
- As Fun is relatively far from the origin we know that Fun is a relatively good discriminator between the brands.
- As Yahoo was very far from the origin, and associated with Fun, we can conclude that Yahoo and Fun were closely associated in 2012 (remember, correspondence analysis focuses on relativities; in 2012 Yahoo's Fun score was half of Google's).
- From 2012 to 2017, Yahoo moved much closer to the origin, which tells us that Yahoo's relative strengths in terms of Fun, Easy-to-Use, and Innovative, have likely declined (and, in reality, they have declined sharply; Google is now more than four times as fun).

The math of correspondence analysis

> This chapter provides a relatively gentle explanation of the mathematics of how correspondence analysis maps are computed. There is no need to read it unless interested.

The data used in this chapter shows the relationship between thoroughness of newspaper readership and education level.[5] It is a *contingency table*, which is to say that each number in the table represents the number of people in each pair of categories. For example, the cell in the top-left corner tells us that 5 people with some primary education glanced at the newspaper. The table shows the data for a sample of 312 people (which is also the sum of the numbers displayed).

	Category of readership		
Level of education	Glance	Fairly thorough	Very thorough
Some primary	5	7	2
Primary completed	18	46	20
Some secondary	19	29	39
Secondary completed	12	40	49
Some tertiary	3	7	16

In the rest of this chapter I refer to the input table as **N**.

The first step in correspondence analysis is to sum up all the values in the table. Let us refer to this as **n**, where:

n = sum(N)

[5] The data in the example comes from Greenacre and Hastie's 1987 paper "The geometric interpretation of correspondence analysis", published in the Journal of the American Statistical Association.

Where practical, the notation and terminology used is from Michael Greenacre's (2016) third edition of Correspondence Analysis in Practice. This excellent book contains many additional calculations for correspondence analysis diagnostics. The only intentional large deviation from Greenacre's terminology relates to the description of the normalizations (see the next chapter).

This chapter draws on: Tim Bock (2011), "Improving the display of correspondence analysis using moonplots", *International Journal of Market Research*.

The next step is to compute a table of proportions, **P**, which involves dividing the input table by the sum (i.e., **n**):

P = N / n

This gives us the following table:

	Glance	Fairly thorough	Very thorough
Some primary	.016	.022	.006
Primary completed	.058	.147	.064
Some secondary	.061	.093	.125
Secondary completed	.038	.128	.157
Some tertiary	.010	.022	.051

Row and column masses

In the language of correspondence analysis, the sums of the rows and columns of the table of proportions are called masses. These are the inputs to lots of different calculations.

The column masses in this example show that Glance, Fairly thorough, and Very *thorough* describe the reading habits of 18.3%, 41.3%, and 40.4% of the sample respectively.

The row masses are Some primary (4.5%), Primary completed (26.9%), Some secondary (27.9%), Secondary completed (32.4%), and Some tertiary (8.3%).

Expected proportions (E)

Referring to the table of proportions, 1.6% (0.016) of people glanced and had some primary education. Is this number big or small? We can compute the value that we would expect to see under the assumption that there is no relationship between education and readership. The proportion that glances at a newspaper is 18.2% and 4.5% have only Some primary education. Thus, if there is no relationship between education and readership, we would expect that 4.5% of 18.2% of people (i.e., 0.008 = 0.8%) have both glanced and have primary education. We can compute the expected proportions of all the cells in the table in the same way, which gives the following table:

	Glance	Fairly thorough	Very thorough
Some primary	.008	.019	.018
Primary completed	.049	.111	.109
Some secondary	.051	.115	.113
Secondary completed	.059	.134	.131
Some tertiary	.015	.034	.034

Residuals (R)

The residuals are computed by subtracting the expected proportions from the observed proportions. Residuals in correspondence analysis have a different role from that which is typical in statistics. Typically, in statistics, the residuals quantify the extent of error in a model. In correspondence analysis, by contrast, the whole focus is on examining the residuals.

The residuals quantify the difference between the observed data and the data we would expect under the assumption that there is no relationship between the row and column categories of the table (i.e., education and readership, in our example).

R = P - E

	Glance	Fairly thorough	Very thorough
Some primary	.008	.004	-.012
Primary completed	.009	.036	-.045
Some secondary	.010	-.022	.012
Secondary completed	-.021	-.006	.026
Some tertiary	-.006	-.012	.018

The biggest residual is -0.045 for Primary completed and Very thorough. That is, the observed proportion of people that only completed primary school and are very thorough is 6.4%, and this is 4.5% lower than the expected proportion of 10.9%, which is computed under the assumption of no relationship between newspaper readership and education. Thus, the tentative conclusion that we can draw from this is that there is a negative association between having completed primary education and reading very thoroughly. That is, people with only primary school education are less likely to read very thoroughly than the average person.

Indexed residuals (I)

Take a look at the top row of the residuals shown in the table above. All of the numbers are close to 0. The obvious explanation for this – that having some primary education is unrelated to reading behavior – is not correct. The real explanation is all the observed proportions (**P**) and the expected proportions (**E**) are small because only 4.6% of the sample had this level of education. This highlights a problem with looking at residuals from a table. By ignoring the number of people in each of the rows and columns, we end up being most likely to find results only in rows and columns with larger totals

(masses). We can solve this problem by dividing the residuals by the expected values, which gives us a table of indexed residuals (**I**).

I = R / E

	Glance	Fairly thorough	Very thorough
Some primary	.95	.21	-.65
Primary completed	.17	.32	-.41
Some secondary	.20	-.19	.11
Secondary completed	-.35	-.04	.20
Some tertiary	-.37	-.35	.52

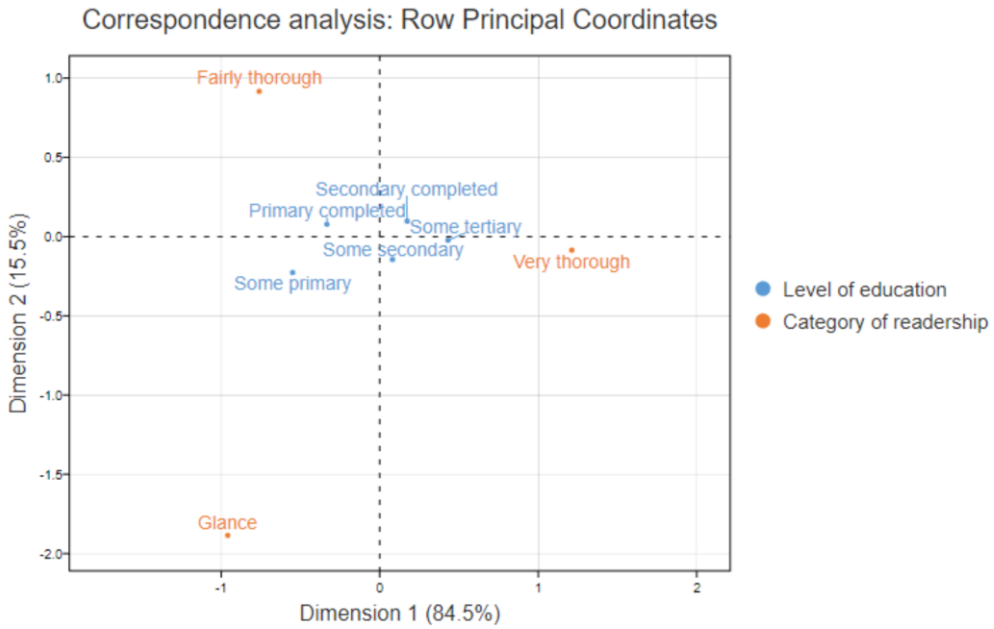
The indexed residuals have a straightforward interpretation. The further the value from zero, the larger the observed proportion relative to the expected proportion. We can now see a clear pattern. The biggest value on the table is the .95 for Some primary and Glance. This tells us that people with some primary education are almost twice as likely to Glance at a newspaper as we would expect if there were no relationship between education and reading. In other words, the observed value is 95% higher than the expected value. Reading along this first row, we see that there is a weaker, but positive, indexed residual of 0.21 for Fairly thorough and Some primary. This tells us that people with some primary education were 21% more likely to be fairly thorough readers that we would expect. And, a score of -.65 for Very thorough, tells us that people with Some primary education were 65% less likely to be Very thorough readers than expected. Reading through all the numbers on the table, the overall pattern is that higher levels of education equate to a more thorough readership.

As we will see later, correspondence analysis is a technique designed for visualizing these indexed values.

Reconstituting indexed residuals from a map

The chart below is a correspondence analysis with the coordinates computed using row principal normalization (its computation is discussed later in this eBook). We can work backward from this map

and compute the indexed residuals, in much the same way that we can recreate orange juice from orange juice concentrate. Some primary has coordinates of (-.55, -.23) and Glance's coordinates are (-.96, -1.89). We can compute the indexed value by multiplying together the two x coordinates and the two y coordinates and summing them up. Thus we have -.55*-.96 + -.23 * -1.89 = .53 + .44 = .97. Taking rounding errors into account, this is identical to the value of .95 shown in the table above.

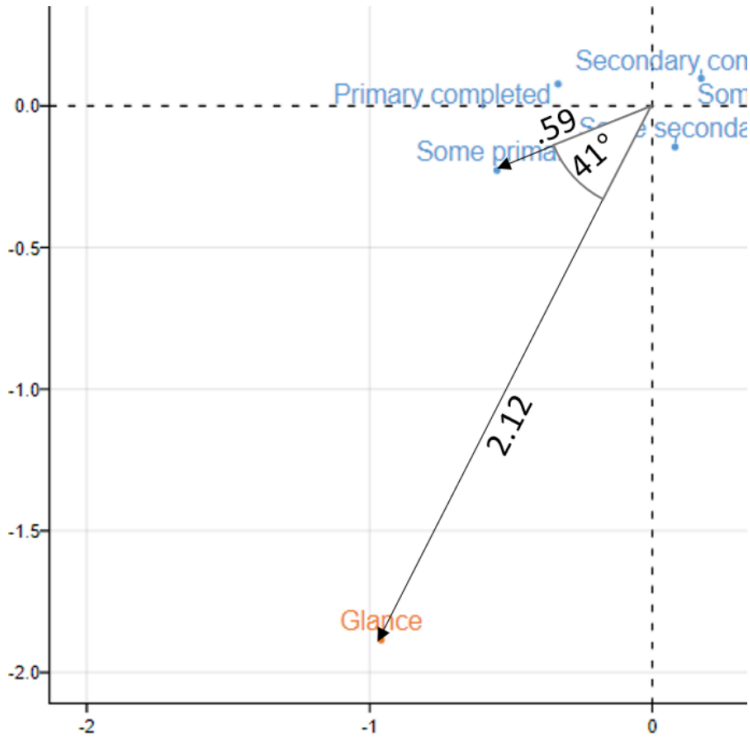


Indexed residual example

Unless you have studied some linear algebra, there is a good chance that this calculation, known as the dot product (or a scalar product or inner product), is not intuitive. Fortunately, it can be computed it in a different way that makes it more intuitive.

To compute the indexed residual for a couple of points, we start by measuring the distance between each of the points and the origin (see the image below). In the case of *Some primary*, the distance is .59. Then, we compute the distance for Glance, which is 2.12. Then we compute the angle formed when we draw lines from each of the points to the origin. This is 41 degrees. Lastly, we multiply

together each of these distances with the cosine of the angle. This gives us .59*2.12*cos(41°) = .59*2.12*.76 = .94. Once rounding errors are taken into account, is the same as the correct value of .95.

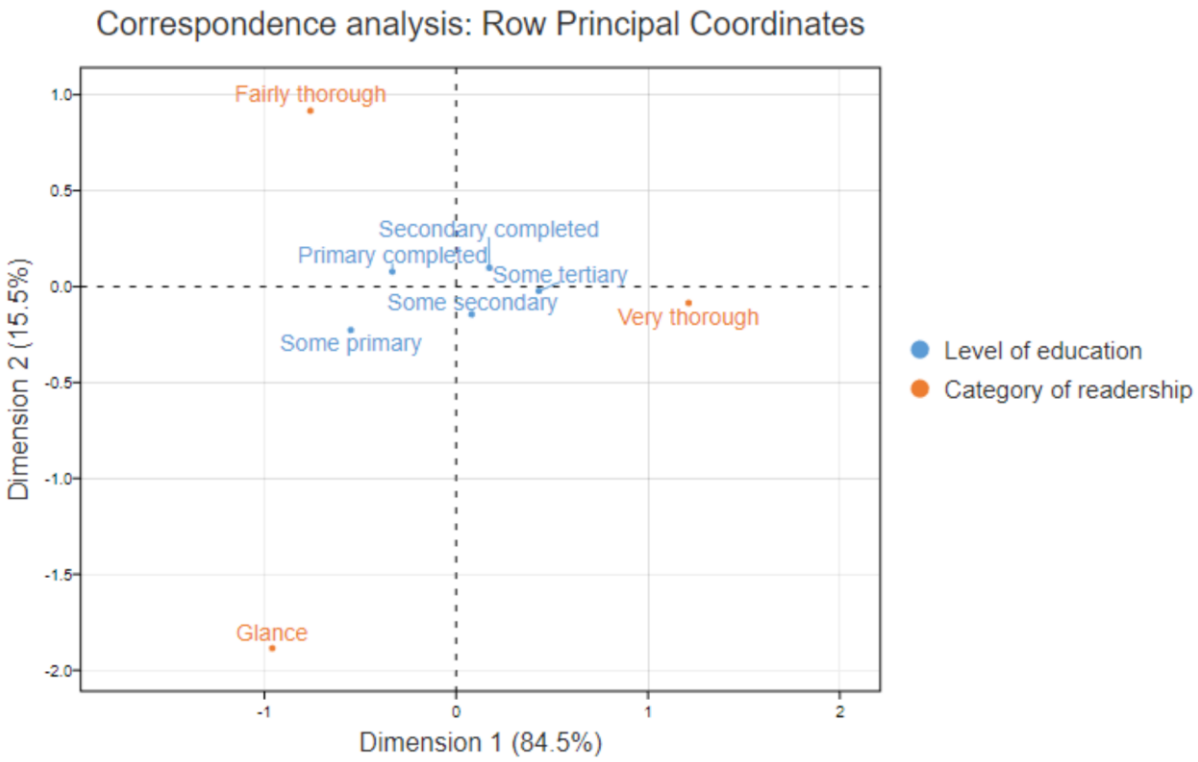


Now, perhaps this new formula looks no simpler than the dot product, but if you look at it a bit closer, it becomes straightforward. The first two parts of the formula are the distance of each point from the origin (i.e., the (0,0) coordinate). Thus, all else being equal, the further the point is from the origin, the stronger the associations between that point and the other points on the map. So, looking at the top, we can see that the column category of Glance is the one which is most discriminating in terms of the readership categories.

The second part to the interpretation, which will likely bring you back to high school, is the meaning of the cosine. If two points are in exactly the same direction from the origin (i.e., they are on the same line), the cosine of the angle is 1. The bigger the angle, the smaller the cosine, until we get to a right-angle (90° or 270°), at which point the cosine is 0. And, when the lines are going in exactly opposite directions (i.e., so the line between the two points goes through the origin), the cosine of the angle is -1. So, when you have a small angle from the lines connecting the points to the origin, the association is relatively strong (i.e., a positive indexed residual). When there is a right angle there is no association (i.e., no residual). When there is a wide angle, a negative residual is an outcome.

Putting all this together allows us to work out the following things from the row principal correspondence analysis map above, which I have reproduced below to limit scrolling:

- People with only Primary completed are relatively unlikely to be Very thorough.
- Those with Some primary are more likely to Glance.
- People with Primary completed are more likely to be Fairly thorough.
- The more education somebody has, the more likely they are to be Very thorough.



Reconstituting residuals from bigger tables

The chart above shows the percentages of variance explained by the map in the labels for the x and y axes. (Their computation is described below.) They indicate how much of the variation in the indexed

residuals is explained by the horizontal and vertical coordinates. As these add up to 100%, we can perfectly reconstitute the indexed residuals from the data. For most tables, however, they add up to less than 100%. This means that there is some degree of information missing from the map. This is not unlike reconstituted orange juice, which falls short of fresh orange juice.

Standardized residuals

The previous few sections describe the relationship between the coordinates on the map and the indexed residuals. This section describes how to compute the coordinates to use when plotting the position of labels on a map. The first step is to compute the standard residuals:

Z = I * sqrt(E)

Standardized residuals are a cool type of statistic in their own right. However, in correspondence analysis, they are computed as an input to a singular value decomposition. The way to think about them in the context of correspondence analysis is that by multiplying the indexed residuals by the square root of the expected proportions, we are essentially weighting the analysis, such that cells with a higher expected value are given a higher weight in the analysis. As the expected values are often related to the sample size, this weighting means that smaller cells on the table, for which the sampling error will be larger, are down-weighted. In other words, this weighting makes correspondence analysis relatively robust to outliers caused by sampling error, when the table being analyzed is a contingency table.

Singular values, eigenvalues, and variance explained

Then, we analyze **Z** using the singular value decomposition*:*

svd(Z)

The singular value decomposition (SVD), which is described in relatively simple terms in An Intuitive Explanation of the Singular Value Decomposition (SVD): A Tutorial in R, converts a table of numbers into three outputs:

- A vector, **d**, which contains the *singular values*.
- A matrix (table) **u** which contains the *left singular vectors*.
- A matrix **v** with the right singular vectors.

```
$d
[1] 2.652708e-01 1.135421e-01 4.212255e-17

$u
                          [,1]        [,2]        [,3]
Some primary        -0.4386666 -0.42375592 -0.08372409
Primary completed   -0.6516462  0.35501142 -0.62648004
Some secondary       0.1603076 -0.67246939 -0.42440653
Secondary completed  0.3711005  0.48847409 -0.27355249
Some tertiary        0.4685240 -0.05979793 -0.58784451

$v
                       [,1]        [,2]       [,3]
Glance          -0.4097795 -0.80584644 -0.4274252
Fairly thorough -0.4887795  0.58960413 -0.6430097
Very thorough    0.7701788 -0.05457549 -0.6354889
```

The left singular vectors correspond to the categories in the rows of the table and the right singular vectors correspond to the columns. Each of the singular values, and the corresponding vectors (i.e., columns of u and v), correspond to a dimension. As we will see, the coordinates used to plot row and column categories are derived from the first two dimensions.

Squared singular values are known as *eigenvalues*. The eigenvalues in our example are .0704, .0129, and .0000.

Each of these eigenvalues is proportional to the amount of variance explained by the columns. By summing them up and expressing them as a proportion, we compute that the first dimension of our correspondence analysis explains 84.5% of the variance in the data and the second 15.5%, which are the numbers shown in x and y labels of the scatter plot shown earlier. The third dimension explains 0.0% of the variance, so we can ignore it entirely. This is why we are able to perfectly reconstitute the indexed residuals from the correspondence analysis plot.

Standard coordinates

As mentioned, we have weighted the indexed residuals prior to performing the SVD. So, in order to get coordinates that represent the indexed residuals, we now need to unweight the SVD's outputs. We do this by dividing each row of the left singular vectors by the square root of the row masses (defined near the beginning of this chapter). This gives us the standard coordinates of the rows:

	[,1]	[,2]	[,3]
Some primary	-2.07	-2.00	-.40
Primary completed	-1.26	.68	-1.21
Some secondary	.30	-1.27	-.80
Secondary completed	.65	.86	-.48
Some tertiary	1.62	-.21	-2.04

We do the same process for the right singular vectors, except we use the column masses. This gives us the standard coordinates of the columns, shown below. These are the coordinates that have been used to plot the column categories on the maps in this chapter.

	[,1]	[,2]	[,3]
Glance	-.96	-1.89	-1.00
Fairly thorough	-.76	.92	-1.00
Very thorough	1.21	-.09	-1.00

Principal coordinates

The principal coordinates are the standard coordinates multiplied by the corresponding singular value. The positions of the row categories shown on the earlier plots are these principal coordinates. The principal coordinates for the education levels (rows) are shown in the table below.

	[,1]	[,2]	[,3]
Some primary	-.55	-.23	-.00
Primary compl-eted	-.33	.08	-.00
Some secondary	.08	-.14	-.00
Secondary completed	.17	.10	-.00
Some tertiary	.43	-.02	-.00

The principal coordinates represent the distance between the *row profiles* of the original table. The row profiles are shown in the table below. They are input data table (**N**) divided by the row totals. Outside of correspondence analysis, they are more commonly referred to as the row percentages of the contingency table. The more similar two rows' principal coordinates, the more similar their row profiles. More precisely, when we plot the principal coordinates, the distances between the points are *chi-square* distances. These are the distances between the rows weighted based on the column masses.

	Glance	Fairly thorough	Very tho-rough
Some primary	.36	.50	.14
Primary completed	.21	.55	.24
Some secon-dary	.22	.33	.45
Secondary completed	.12	.40	.49
Some tertiary	.12	.27	.62

The principal coordinates for the columns are computed in the same way.

In the row principal plot shown earlier, the row categories' positions are the principal coordinates. The column categories are plotted based on the standard coordinates. This means that it is valid to compare row categories based on their proximity to each other. It is also valid to understand the relationship between the row and column coordinates based on their dot products. But, it is not valid to compare the column points based on their position. This is discussed in more detail in the next chapter.

Quality

We have already looked at one metric of the quality of a correspondence analysis: the proportion of the variance explained. We can also compute the quality of the correspondence analysis for each of the points on a map. Recall that the further a point is from the origin, the more that point is explained by the correspondence analysis. When we square the principal coordinates and express these as row proportions, we get measures of the quality of each dimension for each point. Sometimes these are referred to as the squared correlations and squared cosines.

The quality of the map for a particular category (i.e., label that appears on the visualizations) is usually defined as the sum of the scores of the category for the two dimensions that are plotted. In our example, these all add up to 100%. However, in a more typical example, with a bigger data table, it is rare for categories to have 100% of their variance explained.[6]

	[,1]	[,2]	[,3]
Some primary	.854	.146	.000
Primary completed	.948	.052	.000
Some secondary	.237	.763	.000
Secondary completed	.759	.241	.000
Some tertiary	.997	.003	.000
Glance	.585	.415	.000
Fairly thorough	.790	.210	.000
Very thorough	.999	.001	.000

[6] Where a table has less than or equal to three rows or columns it is guaranteed that 100% of the variance is explained by two dimensions.

Other types of data

Although the explanation of the math of correspondence analysis in this chapter is closely tied to the notion that the data is a contingency table, it is routine to use the same mathematics to analyze other types of tables.

Normalization and scaling

Correspondence analysis is useful because it compresses information, replacing a detailed table with an easier-to-interpret visualization. As is the case with most simplifications, some information is lost.

A decision made when conducting correspondence analysis is the choice of *normalization,* which determines which information is lost. Most correspondence analyses plots are misleading in at least three different ways, but the choice of normalization can increase this to five.

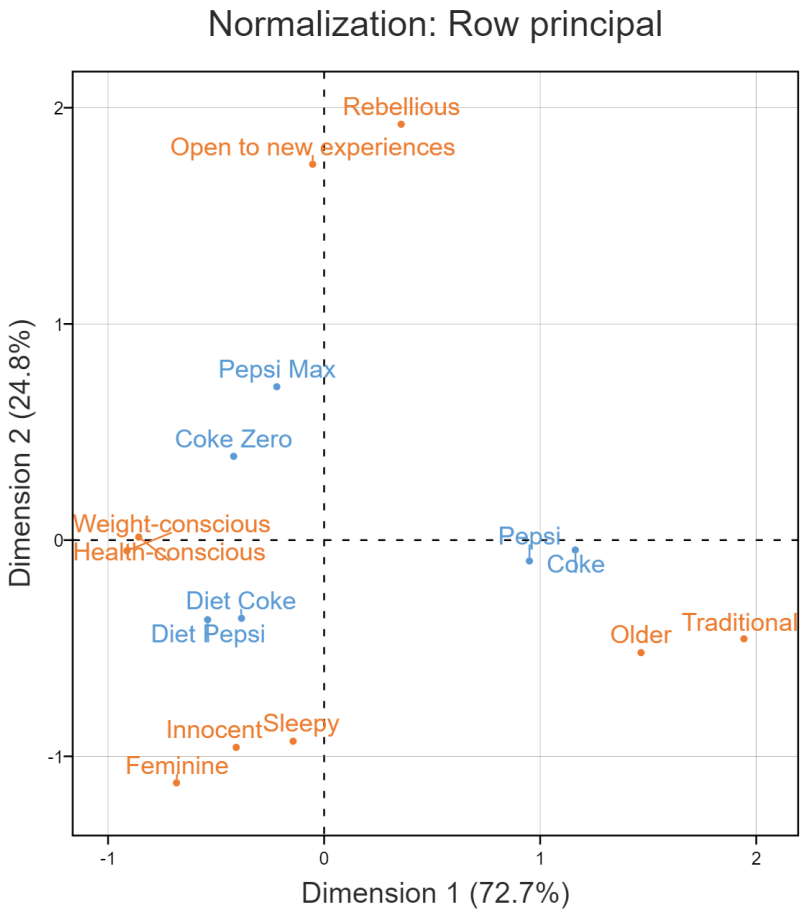
This chapter provides an overview of the main normalization options, explains how to interpret the resulting maps, provides a technical explanation of the normalizations, and recommendations about normalization for different situations.

The table below lists the main *normalizations* used in correspondence analysis, the key concepts and terminology used. By default, most programs show **Principal**.

Normalization	Other names	Definition of row coordinates	Definition of column coordinates	How to interpret relationships between row coordinates	How to interpret relationships between column coordinates	How to interpret relationships between row and column categories
Standard	Symmetrical[7]	Standard	Standard	The vertical distances are exaggerated	The vertical distances are exaggerated	No straightforward interpretation
Row principal	Row, Row asymmetric, Asymmetric map of the rows, Row-metric-preserving	Principal	Standard	Proximity	The vertical distances are exaggerated	Dot product
Row principal (scaled)		Principal	Standard × first eigenvalue	Proximity	The vertical distances are exaggerated	Proportional dot product
Column principal (scaled)	Column, Column asymmetric, Asymmetric map of the columns, Column-metric-preserving	Standard × first eigenvalue	Principal	The vertical distances are exaggerated	Proximity	Proportional dot product
Column principal		Standard	Principal	The vertical distances are exaggerated	Proximity	Dot product
Principal	Symmetric map, French scaling, Benzécri scaling, Canonical, Configuration Plot	Principal	Principal	Proximity	Proximity	No straightforward interpretation
Symmetrical (1/2)	Symmetrical, Symmetric, Canonical scaling	Standard × sqrt(singular values)	Standard × sqrt(singular values)	The vertical distances are somewhat exaggerated	The vertical distances are somewhat exaggerated	Dot product

[7] There is no commonly-agreed upon meaning of the term "symmetric(al)" when applied to correspondence analysis normalization. For example, perhaps the most widely used program for correspondence analysis, SPSS Statistics, uses a meaning that is completely different from that of the most widely read author on the topic, Michael Greenacre. For this reason, the term is not used in the rest of this eBook.

The first requirement for correct interpretation of correspondence analysis is a scatterplot with an aspect ratio of 1, which is the technical way of saying that the physical distance on a plot between values on the x-axis and y-axis need to be the same. If you look at the plot below, you will see that the distance between 0 and 1 on the *x*-axis (horizontal) is the same as the on the *y*-axis (vertical), so this basic hurdle has been passed. But, if you are viewing correspondence analysis in general-purpose charting tools, such as Excel or `ggplot`, be careful, as they will not, by default, respect the aspect ratio, which will make the plots misleading (see the example in **5. Proximity between row labels probably indicates similarity (if properly normalized)**).



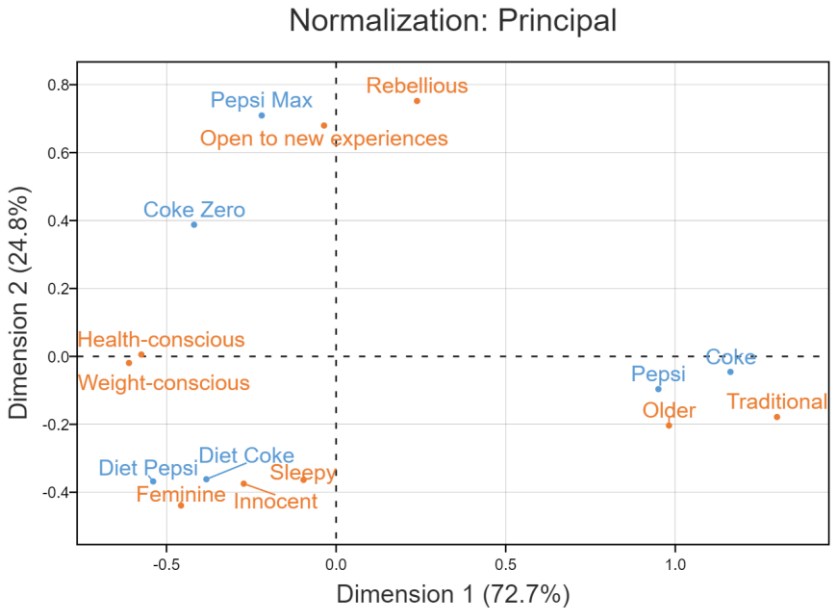
As mentioned at the beginning of this chapter, most standard correspondence analysis plots are misleading in at least three ways.

The first way is that they only show relativities. For example, the plot above suggests that Pepsi and Coke (which were rows in the table) are both associated with Traditional or Older (columns). However, there is no way to conclude from this map which brand has the highest score on any attribute. In the case of maps showing *brand associations* (as in most of the examples in this eBook),

it is quite common to have a leading brand with the highest score on all the attributes; the key when interpreting is to remember that the map only shows relativities.

The second general way that correspondence analysis maps mislead relates to the variance explained. If you add up the percentages in the x and y axis labels above, you will see that they add up to 97.5%. So, 2.5% of the variance in the data is not explained. This is not much. But, the percentage can be much higher. The higher the percentage, the more misleading the plot. And, of course, it is possible that the two dimensions explain 100% of the variance, as is illustrated in the previous chapter.

The map above is misleading in a third way. It misrepresents the relationship between the columns. The plot shows that **Weight-conscious** is roughly the same distance apart from **Older** as it is from **Rebellious**. This is a misrepresentation of the data. To correctly interpret the relationship between the row coordinates, we need to remember that the vertical dimension explains only about a third of the variance, so vertical distances for the column coordinates are on this plot are exaggerated. If you look at the plot below, it shows the relationship between the columns properly.



What is the difference between the two plots? The earlier plot uses *row principal normalization*. This means it gets the rows right, but not the columns. The plot above uses principal normalization, which means it gets the rows and columns correct.

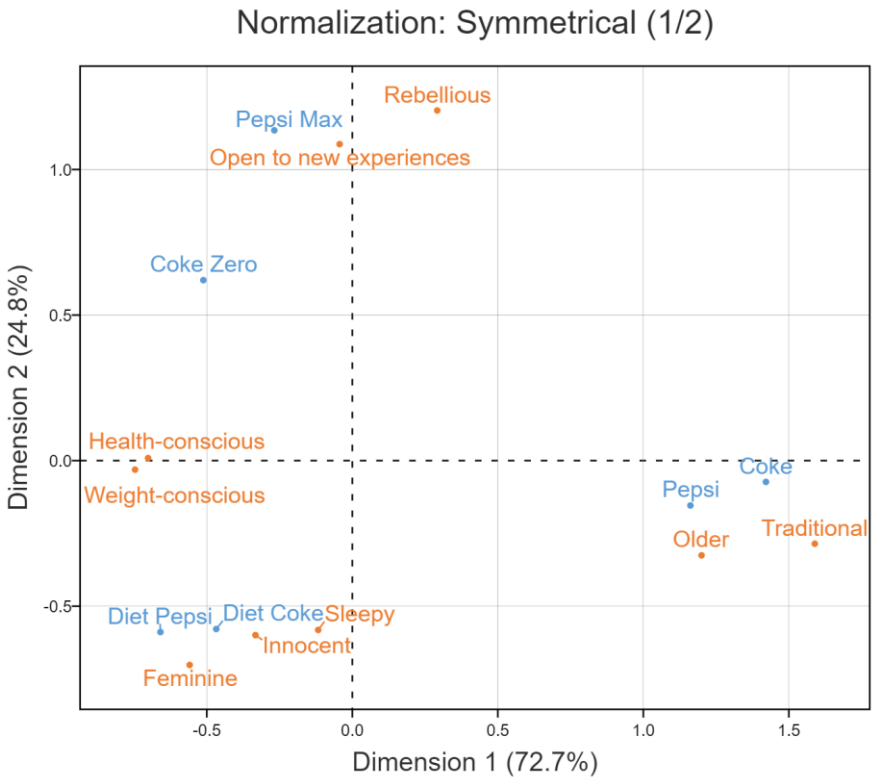
At this stage, it no doubt seems the principal normalization is better. Who would want a map which misrepresented the relationship between the column categories? Unfortunately, the principal normalization comes with its own great limitation.

The principal normalization is great at showing the relationships within the row coordinates, and also within the column coordinates. However, it misrepresents the relationships between the row and the column categories. In the row principal normalization shown above, we can infer the relationship between row and column categories by looking at how far they are from the origin, and also the angle formed by the lines that connect them to the origin.

The misrepresentation of the relationships between the row and column categories can best be described as moderate. Yes, it is not possible to correctly work out all the relationships from the map, even if the map explains 100% of the variance. However, any strong relationships that appear on the map are likely to be correct. This makes the principal normalization a good default normalization. However, in situations where there is a clear focus on the rows, such as when using it to show brand positioning (as in these examples), the row principal normalization is better.

It is also possible to use *column principal normalization*. If I have done a good job in explaining things, you can hopefully work out that this normalization correctly shows the relationships between the rows and the columns,but misrepresents the relationships among the row categories.

The next useful normalization is one that is referred to in Displayr and Q as *symmetric (1/2)* normalization. This normalization, shown below and defined in a bit more detail below, correctly shows the relationship between the row and column coordinates. But, it misrepresents the relationships among the row points, and also among the column points. So, of all the normalization we have seen so far, it is the one that misrepresents the data in the most ways. However, it does have an advantage. Its degree of misrepresentation is the smallest. That is, while the row normalization misrepresents the column coordinates by quite a large amount, the symmetric 1/2 misrepresents them by a smaller amount. Similarly, while the column normalization misrepresents the row coordinates by a large amount, the plot below does so by a smaller amount.

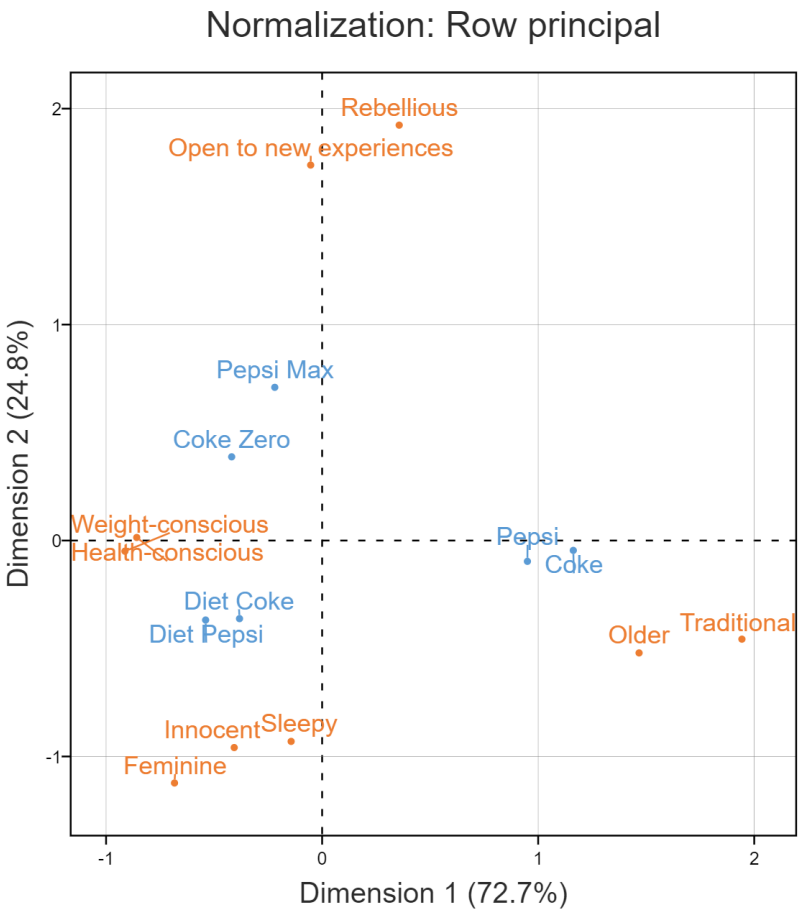


The consequence of this is that if in a situation where the main interest is in the relationships between the row and column coordinates, and there is no clear way of knowing whether to choose row or column principal normalization, this approach is the best one.

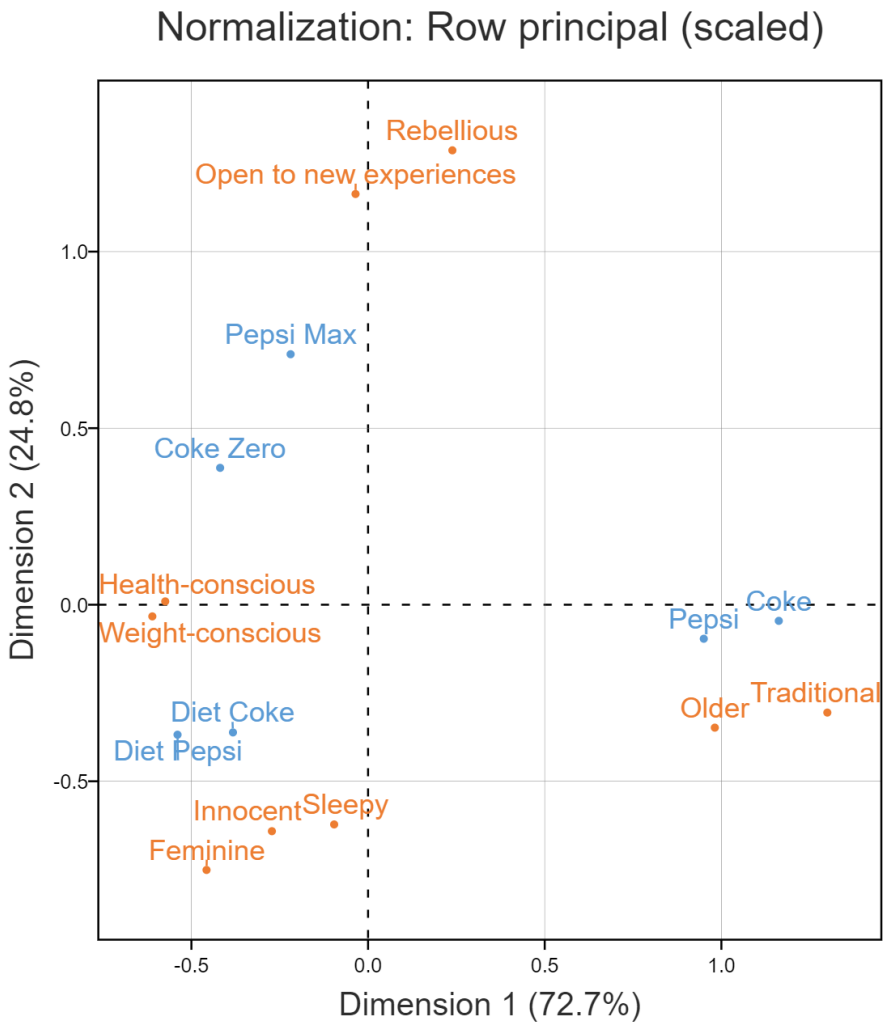
Row principal normalization is probably the best choice of normalization used so far, making sure that the key categories to be compared are shown in the rows of the table being analyzed. This is good practice for two reasons:

- In most situations, either the rows or columns are likely to be the focus. In the examples in this eBook, for example, where the table shows brands, in most cases the brands are shown in the rows (the exception to this is when the resulting table is too wide to be viewed).
- Due to the ease with which people can misunderstand the normalizations, it is useful to train viewers to learn one of them, so always using row principal is likely to lead to fewer misunderstandings than if sometimes using column and other times using row normalization.

The row principal normalization is, however, imperfect. Below I have repeated the row principal plot from the beginning of the chapter. A practical problem with this normalization is that the row categories tend to cluster in the middle of the map and the column categories at the periphery. Sometimes this can make it impossible to read the row categories, as they are all overlapping.



A straightforward improvement on the row principal normalization is to scale the column coordinates on the same scale as the x-axis of the row coordinates. This results in what Q and Displayr refer to as *row principal (scaled) normalization*. As I discuss in the next section, this is an improvement without cost.



A technical explanation of the different normalizations

Below are the core numerical outputs of a correspondence analysis of the data used so far in this chapter. The first row shows the singular values. The remaining rows show the standard coordinates for the rows (brands) and columns (attributes). (Refer to the previous chapter for a detailed explanation about what these are and how they are computed.)

Singular value	0.669	0.391	0.104	0.055	0.034
Coke	1.74	-.12	.61	-.67	-1.14
Diet Coke	-.57	-.93	1.25	1.10	.12
Coke Zero	-.63	.99	.72	-1.33	.99
Pepsi	1.42	-.25	-1.17	.75	2.00
Diet Pepsi	-.81	-.94	-1.25	-.64	-.54
Pepsi Max	-.33	1.81	-.63	1.26	-.89
Feminine	-.68	-1.12	.02	1.02	-.49
Health-conscious	-.86	.01	.96	-.59	.76
Innocent	-.41	-.96	-2.43	-1.64	-1.48
Older	1.47	-.52	.67	.55	-.46
Open to new experiences	-.05	1.74	-.44	-1.05	.79
Rebellious	.36	1.92	-.71	1.75	-1.06
Sleepy	-.14	-.93	-1.69	1.43	2.72
Traditional	1.94	-.46	.38	-.76	.23
Weight-conscious	-.91	-.05	.82	-.02	-.54

The row principal normalization is computed by multiplying the position coordinates of each of the row categories from the original table (i.e., Coke through Pepsi Max) by the corresponding singular values. The first two dimensions are then plotted. For example, Coke Zero's coordinate on the x-axis is .669*-0.63 = -.42, and its position on the y-axis is .391*.99 = .39. As mentioned, if the two dimensions explain all the variance in the data, then the positions of Coke Zero relative to all the other brands on the map is correct.

Expressing these calculations as formulas, we have:

x for a row = Singular value 1 * Standard Coordinate 1

and

y for a row = Singular value 2 * Standard Coordinate 2

For the column categories, we plot the standard coordinates:

x for a column = Standard Coordinate 1

y for a column = Standard Coordinate 2

This simpler formula is not correct. By ignoring the singular values, these coordinates misrepresent the scale. However, the reason for this “mistake” is that the dot product of these coordinates is meaningful. As described in the previous chapter, correspondence analysis allows us to understand the relationships between rows and column categories, where this relationship is formally quantified as the indexed residuals, where:

Indexed residual for x and y = x for row * x for column + y for row * y for column

If you substitute in the earlier formulas this gives us:

Indexed residual for x and y = Singular value 1 * Standard Coordinate 1 * Standard Coordinate 1 + Singular value 2 * Standard Coordinate 2 * Standard Coordinate 2

When we use the principal normalization, this means we use the principal coordinates for both the row and column categories, which changes the formula to (Singular value 1)² * Standard Coordinate 1 * Standard Coordinate 1 + (Singular value 2)² * Standard Coordinate 2 * Standard Coordinate 2. As you can see, this puts the singular values in twice, and so no longer correctly computes the indexed values.

The symmetric (1/2) normalization computes the coordinates for x and y for both row and column coordinates using $\sqrt{\text{singular value}}$ * Standard Coordinate. As the principal coordinates, which multiply by the singular values rather than their square roots, are correct, it follows that this normalization is neither correct for within row comparisons nor for within column comparisons. Nevertheless, its degree of error is lower than standard coordinates. The indexed residuals are correctly computed because $\sqrt{\text{singular value}} * \sqrt{\text{singular value}}$ = Singular value.

The row principal (scaled) normalization uses:

x for a column = Singular value 1 * Standard Coordinate 1

y for a column = Singular value 1 * Standard Coordinate 2

That is, it uses the first singular value for each of the two coordinates. This has the effect of contracting the scatter of the column coordinates on the map, but makes no change to their relativities (i.e., they remain wrong, as they ignore the reality that the y dimension explains less variation). This normalization also changes the indexed residual, so that rather than the dot product being exactly equal to the indexed residual when the plot explains 100% of the variance, instead the dot product becomes proportional to the indexed residual. Changing from an equality to a proportionality has no

practical implication of any kind, as relationships between the row and column categories are only ever interpreted from correspondence analysis as relativities. This is why the scaling of row principal is generally appropriate.

Column principal (scaled) is the same as row principal (scaled), except that the focus is switched from the columns to the rows.

VISUALIZATION

The standard correspondence analysis visualization, which has appeared throughout the earlier chapters, is variously known a *scatterplot*, a *map*, and a *biplot*. This section reviews some alternative plots and variations of the traditional scatterplot:

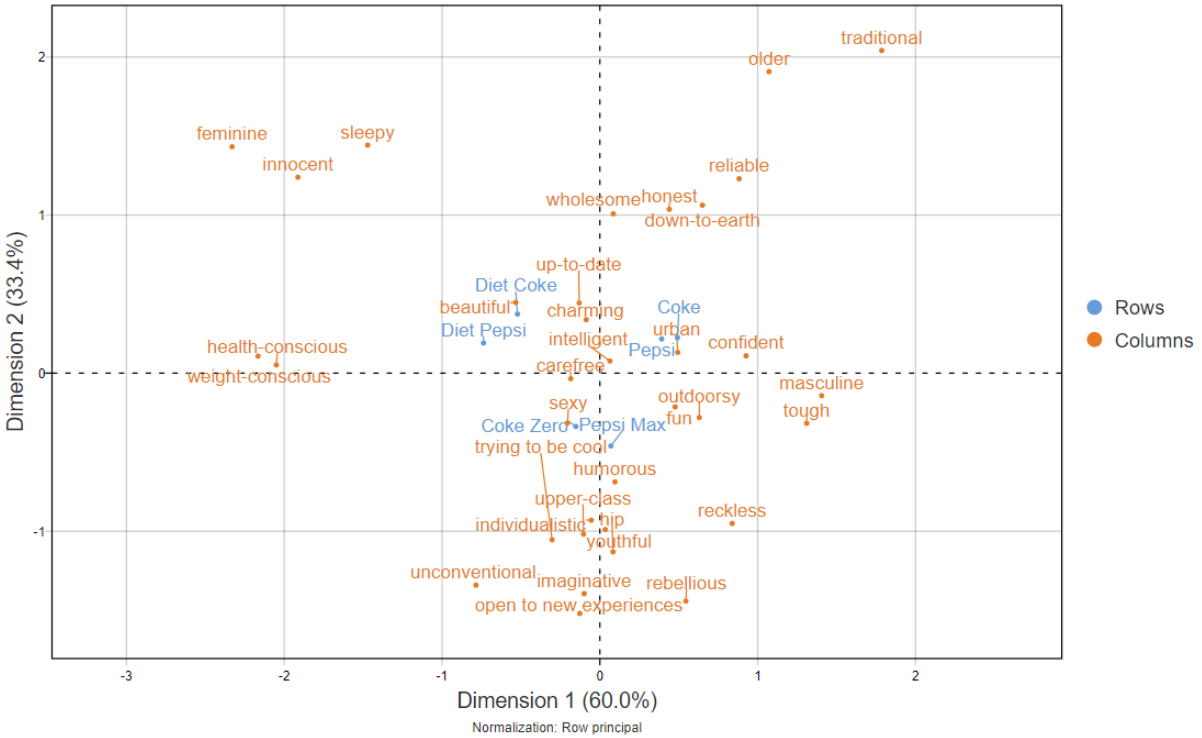
- Moonplots
- Bubble charts
- Tables of studentized residuals
- Heatmaps
- Images
- Trend lines
- 3D visualizations

Moonplots

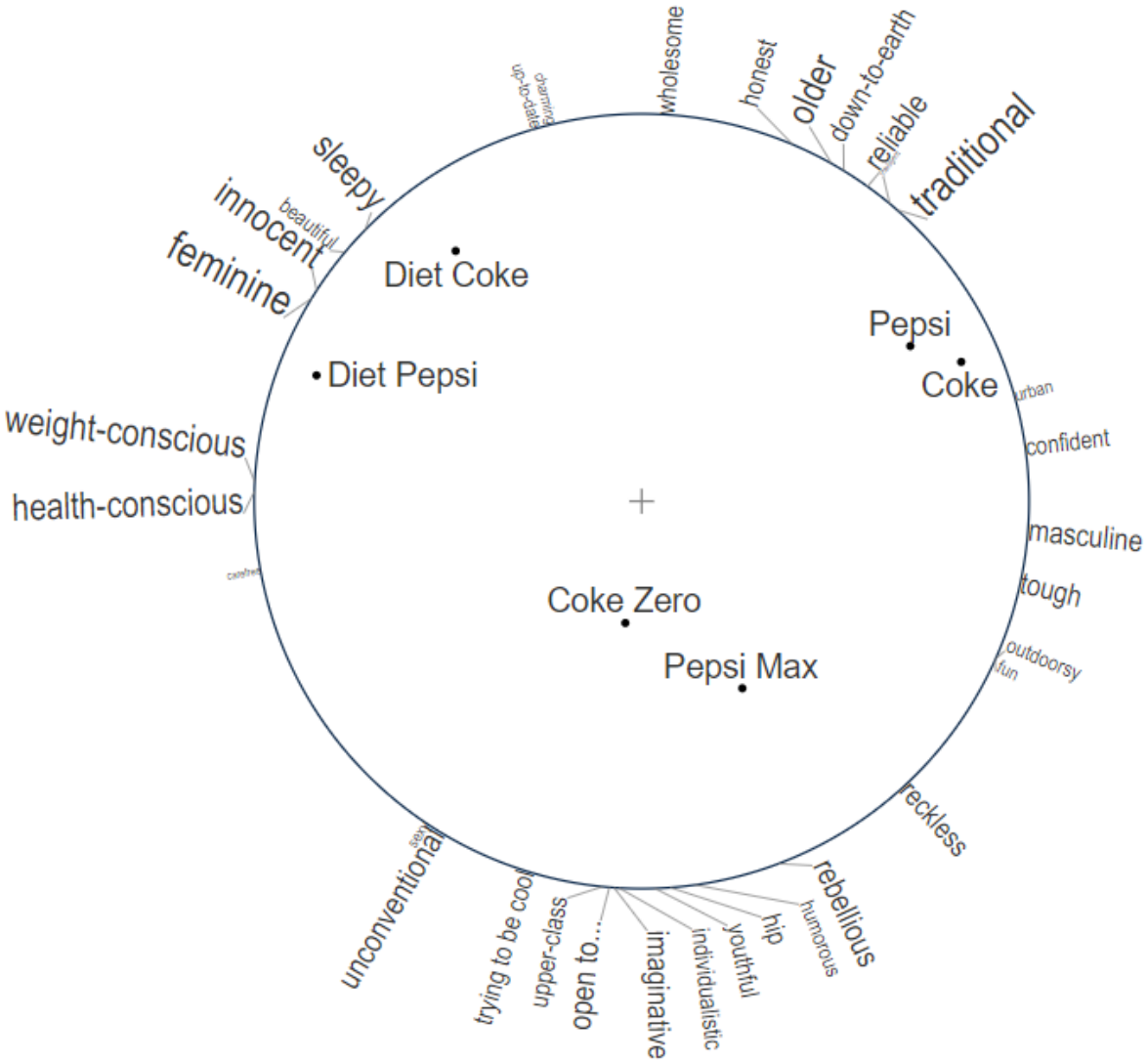
> The traditional correspondence analysis scatterplot can be:
>
> - too messy to use, and
> - easily misinterpreted.
>
> The *moonplot* is an alternative visualization which can overcome these limitations.

Consider the visualization below. It is super-messy. The messiness it not due to a lack of care. The problem is that there are just too many attributes that need to appear in the same space, and no amount of careful arrangement of labels can solve this problem.

A second limitation of this visualization is that less-experienced analysts often misinterpret the center of the visualization, failing to appreciate that when labels are close together in the center of the visualization (0,0), this means that they are unrelated rather than closely related. For example, in the visualization below, the correct interpretation is that there is no relationship between Pepsi and Carefree.



A moonplot of the same data is shown on the next page. The key difference between the moonplot and the traditional visualization (a *scatterplot*) relates to the display of attributes. The scatterplot above plots the attributes in the same space as the brands. By contrast, the moonplot displays the attributes equidistant from the center of the visualization. The font sizes instead communicate the same information which is conveyed in the traditional visualizations by the distance of the attributes to the origin.



Software

A traditional correspondence analysis in Q and Displayr can be converted to a moonplot by setting **Output** to **Moonplot**.

Advantages of the moonplot over traditional brand maps

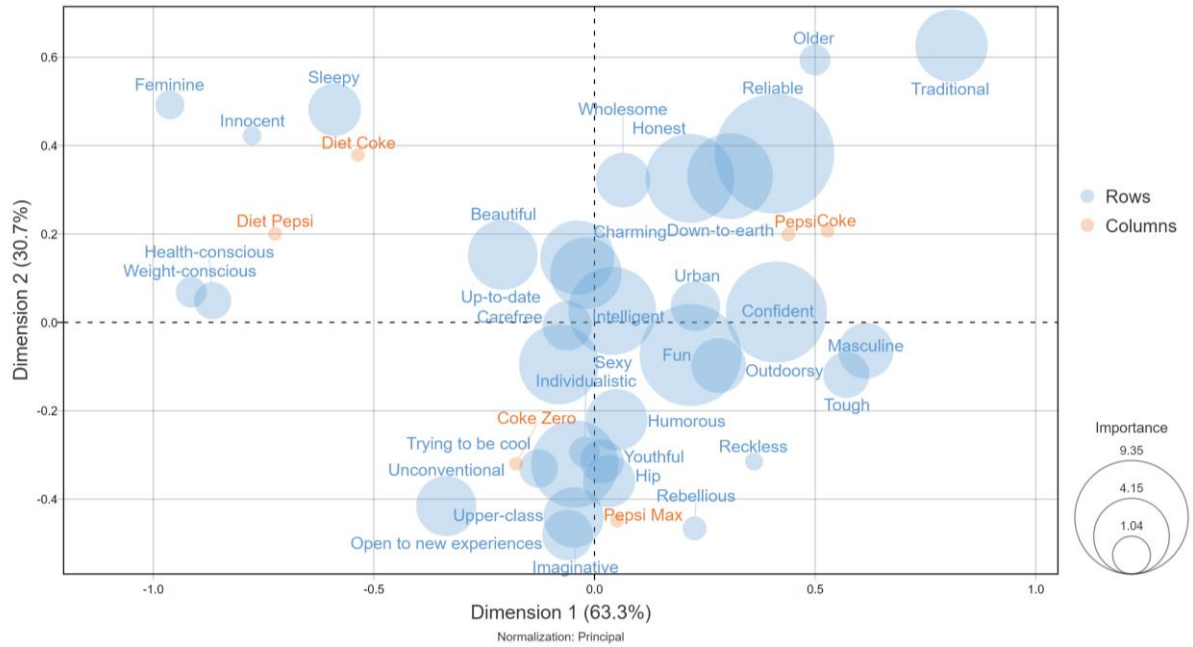
While admittedly a bit uglier than the traditional display, the moonplot visualization has some big advantages:

- It is tidier.
- The tidiness makes it easier to understand the extent to which brands' positions are strong. Coke Zero, and (to a lesser extent), Pepsi Max, are closer to the center of the map than Diet Pepsi and Diet Coke. This means they are less differentiated than the other brands based on the attributes in the study. While an expert can obtain the same conclusion from the traditional map, with the moonplot it is obvious to everyone (novice to expert).
- The varying font sizes make it clear that all attributes are not equal. For example, the small font for Carefree makes it clear that in some sense the attribute is unimportant. To deduce this from the traditional map requires expertise.
- Most importantly, the obvious interpretation of this map is correct in terms of the brand associations. For example, it is clear on this map that Diet Pepsi is associated with Feminine, Innocent, Sleepy, Weight-conscious, and Health-conscious. The user can work this out by glancing at the map, with no need for rulers, protractors, or an understanding of the dot product.

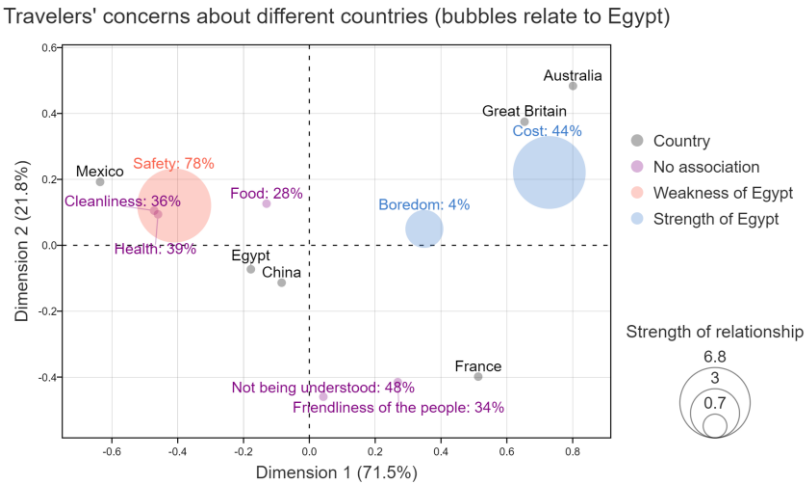
Bubble charts

> A straightforward way to augment a traditional correspondence analysis is to replace the points with bubbles, where the size of the bubbles is used to communicate additional information.

In bubble chart below, *driver analysis* (see our eBook DIY Driver Analysis) has been used to determine the importance of each of the attributes, and this has been represented by the area of each attribute's bubble.



In this next example, which shows concerns about different countries by American travelers, the bubbles have been drawn proportional to the standardized residuals of the concerns for Egypt, with negative relationships shown in pink and positive in blue. Refer to the blog post Customization of Bubble Charts for Correspondence Analysis in Displayr for details about how to create a visualization like this.



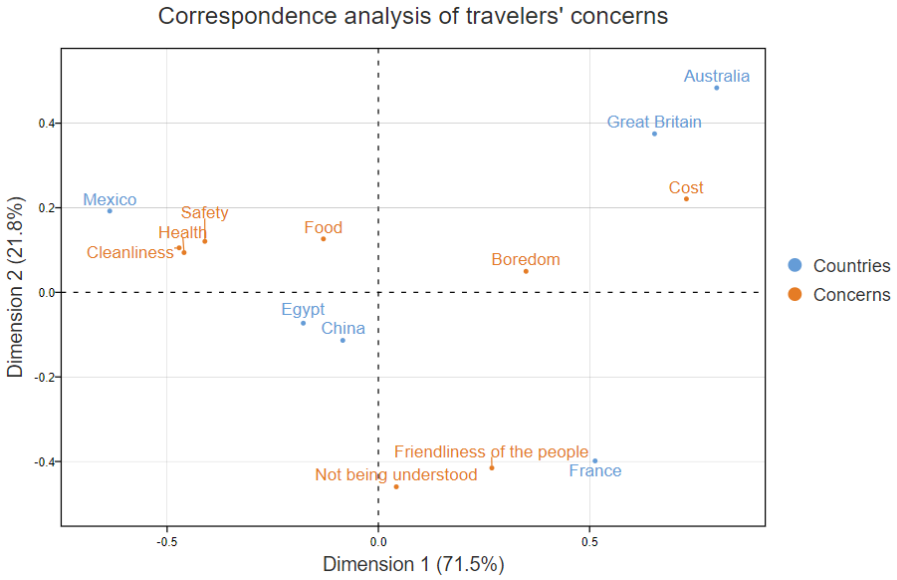
Software

A traditional correspondence analysis in Q and Displayr can be converted to a bubble chart by setting **Output** to **Bubble Chart** and providing a table or R Output that contains the bubble sizes, with matching names to the original table, in the **Bubble sizes** field.

Tables of standardized residuals

> As described in **The math of correspondence analysis**, correspondence analysis is a graphical representation of standardized residuals. It is often helpful to look at the standardized residuals themselves when interpreting correspondence analysis.

In the correspondence analysis below, Egypt and China are near the middle of the map, which may mean that they are undifferentiated. Or, it may mean that they are just less differentiated than the other countries.



The table below shows a type of standardized residual called the *z*-Statistic. It is the same basic idea of the standardized residuals described in **The math of correspondence analysis**, except that they have been scaled to reveal statistical significance, with a value of less than -1.96 or more than 1.96 being statistically significant at the 0.05 level. From this table we can see that Egypt and China are each significantly different from the average on a number of attributes.

z-Statistic	Cleanliness	Health	Safety	Cost	Food	Not being understood	Friendliness of the people	Boredom
Mexico	9.5↑	8.2↑	11.4↑	-16.3↓	.8	-6.7↓	-5.6↓	-1.3
France	-5.3↓	-6.7↓	-8.3↓	6.0↑	-4.1↓	6.9↑	10.6↑	1.9
Great Britain	-2.7↓	-3.3↓	-5.1↓	10.3↑	2.4↑	-4.8↓	-.7	5.7↑
Egypt	-.7	-.1	7.5↑	-7.3↓	-.1	1.0	.4	-2.0↓
Australia	-3.8↓	-3.5↓	-3.9↓	13.8↑	-.2	-5.2↓	-1.4	-.1
China	.3	2.6↑	-3.5↓	-3.7↓	1.3	6.2↑	-2.7↓	-.8

Software

Q and Displayr both automatically display standardized residuals on all tables by the use of color and arrows. A blue arrow indicates that the z-Statistic is positive and statistically significant (by default, the *false discovery rate correction* is set to 0.05). A red arrow indicates negative. The length of the arrows are proportional to the *z*-Statistics.

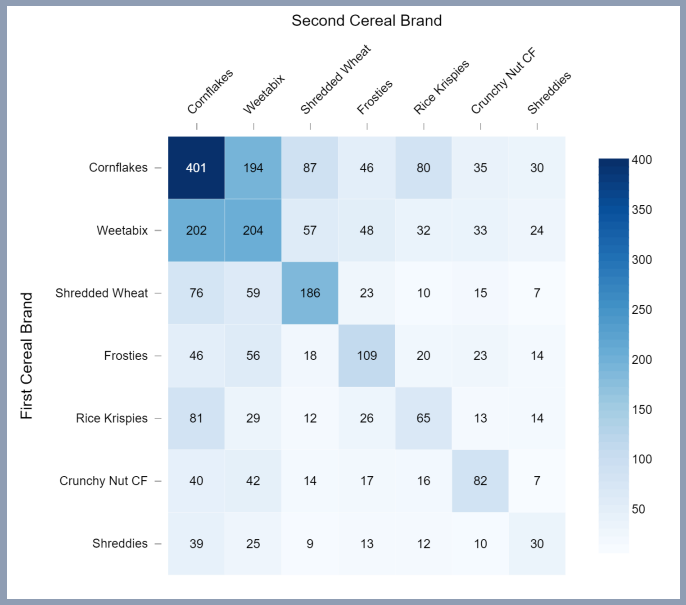
The *z*-Statistics can be viewed directly by clicking on a table and then:

- in Q, right-clicking on the table, and selecting **Statistics – Cells > z-Statistic**
- in Displayr, clicking **Statistics – Cells** (in the object inspector) and selecting **z-Statistic**.

Heatmaps

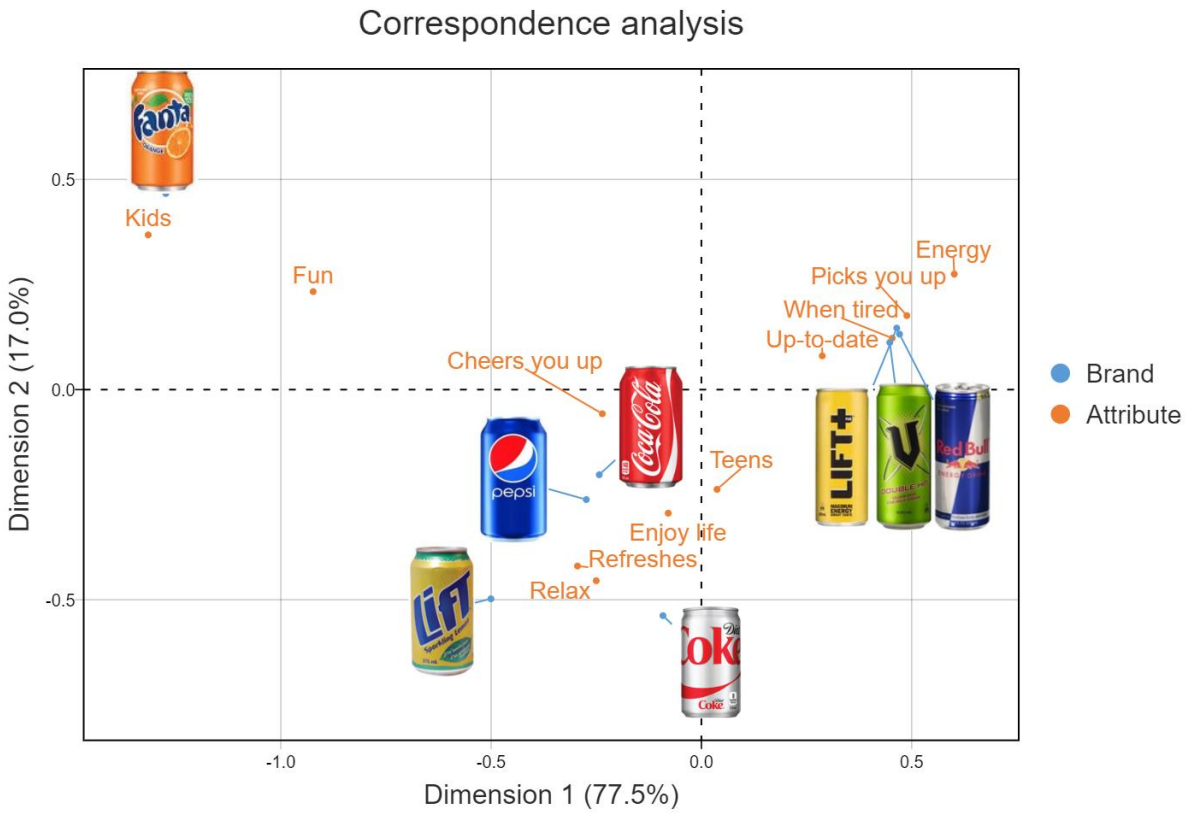
A heatmap of the input table is also a useful way of spotting patterns. The heatmap below, for example, is from the earlier example illustrating correspondence analysis of a square table. The heatmap makes it easy to see the strong association between Cornflakes and Rice Krispies.

In Q: **Create > Charts > Visualization > Heatmap**
In Displayr: **Insert > Visualization > Heatmap**



Images (e.g., logos)

> Replacing labels with images can improve the visualization of a correspondence analysis, as it makes them more attractive and easier to digest.



The visualization above replaces brands with images of cans. In Q and Displayr images can be added to a correspondence visualization by creating the correspondence analysis in the normal way (see **Software**), and then:

- Create logos of similar sizes and put them somewhere on the web. The most straightforward way to do this is to use Dropbox and share the files. Make sure that you check this has worked by pasting the URLs into your browser; if you do not see the logo, something has gone wrong.
- Checking **Use logos for rows**. If you want to instead use logos for the data in the columns, check **Switch rows and columns**.
- Paste your URLs, with commas between them and quotation marks around them, into the **Logos** box and press **Calculate**. The order of the images in the list should match the order of the rows in the table. The list of URLs that I used in the example above looks like this:

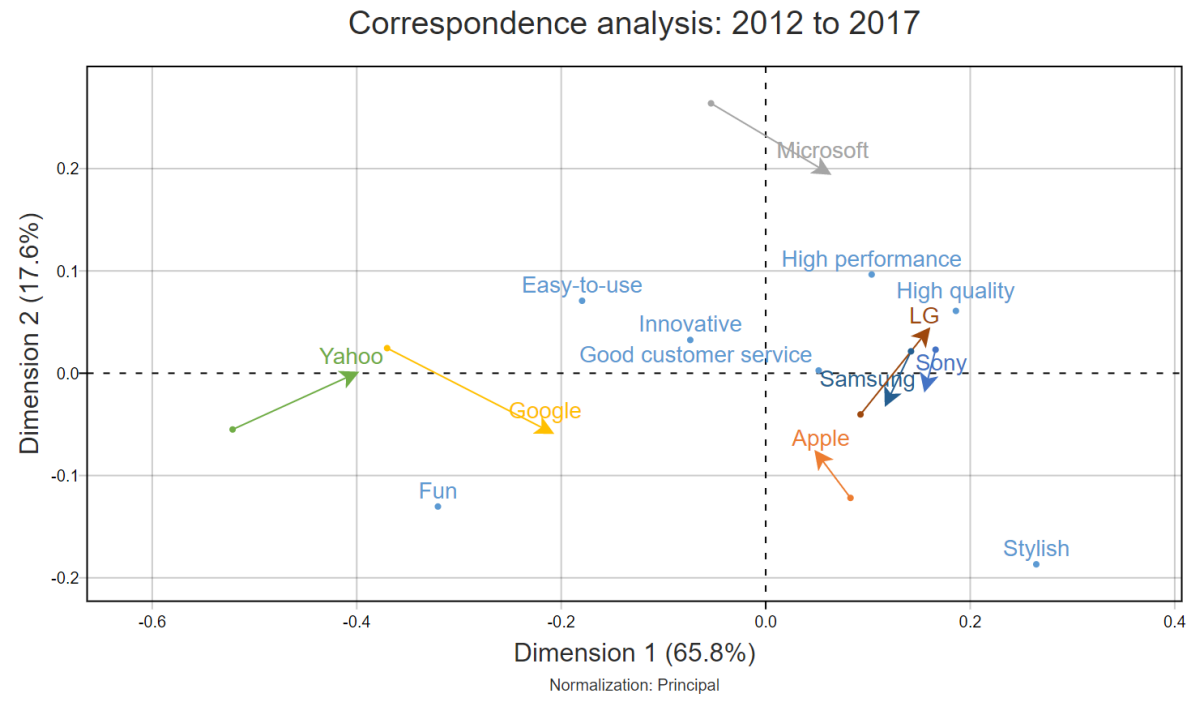
```
"http://docs.displayr.com/images/9/90/Coke.png","http://docs.displayr
.com/images/7/7c/V.jpg","http://docs.displayr.com/images/8/82/RedBull
.jpg","http://docs.displayr.com/images/d/dc/LifePlus.jpg","http://doc
s.displayr.com/images/0/09/DietCoke.png","http://docs.displayr.com/im
```

ages/d/da/Fanta.jpg","http://docs.displayr.com/images/e/e8/Lift.png",
"http://docs.displayr.com/images/5/5e/Pepsi.jpg"

Trend

When using correspondence analysis to compare data over time (see **Correspondence analysis of multiple tables**), it can make things easier to understand if the differences are shown by trend lines, as shown below.

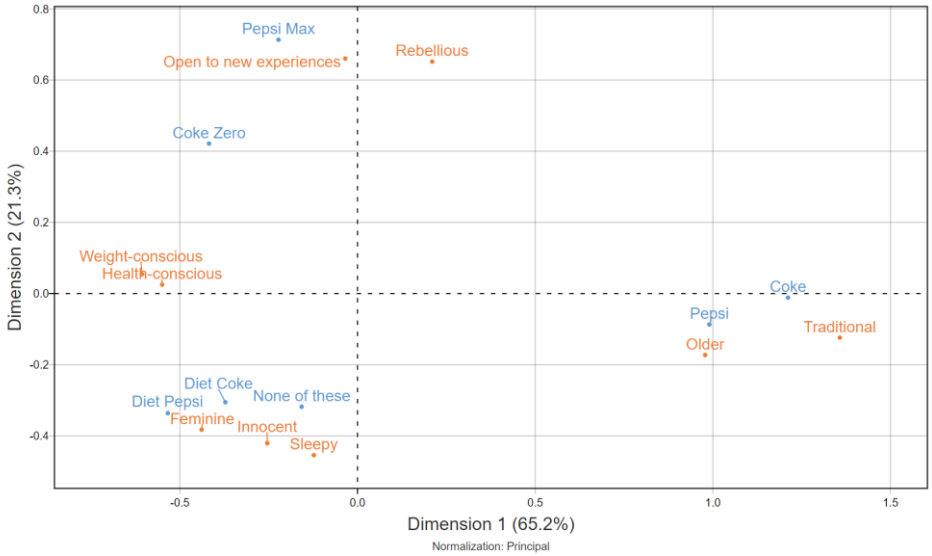
In Q and Displayr, this is achieved by checking the **Trend lines** option (which only appears when you have multiple tables selected).



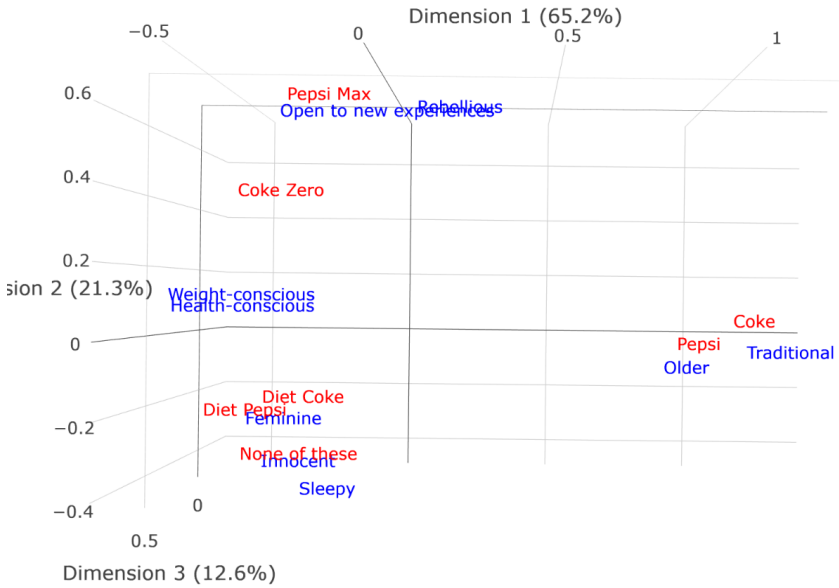
3D visualizations

> Sometimes it can be useful to create 3D scatterplots. This is typically most useful when checking solutions, rather than when presenting them to stakeholders.

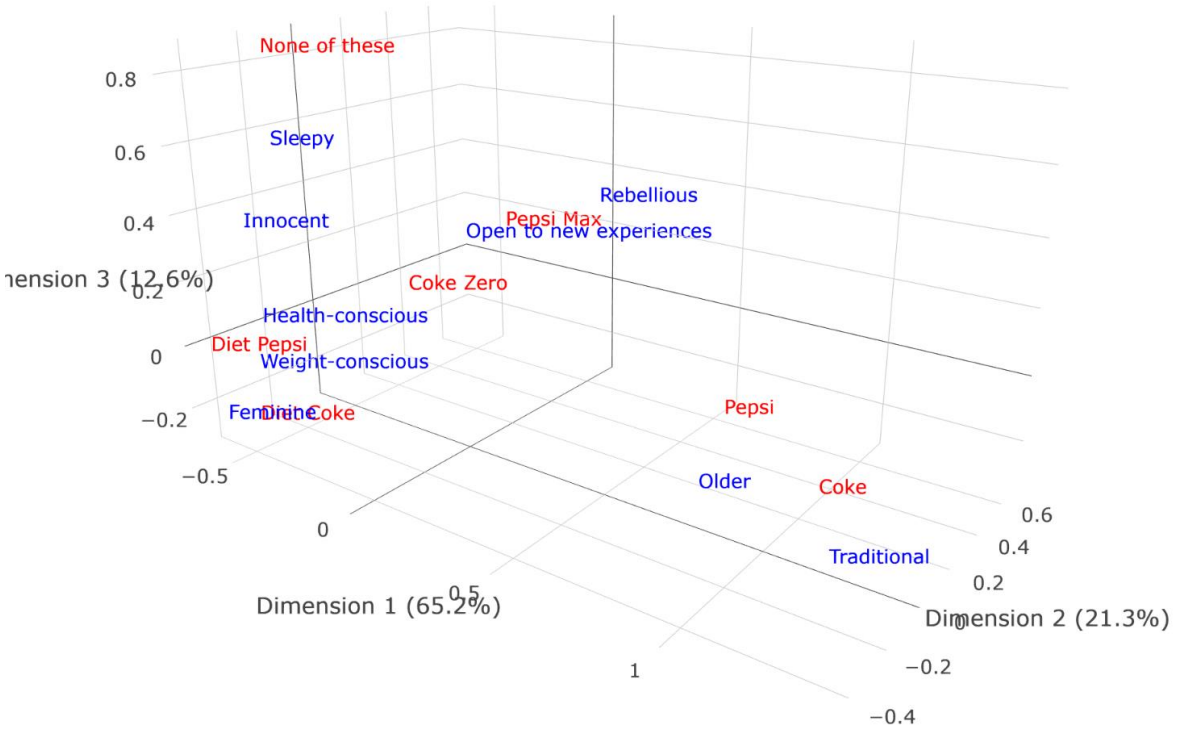
The visualization below is the standard 2D scatterplot visualization from correspondence analysis. It explains 86% of the variance. This leads to the question: is the unexplained 14% interesting?



The visualization below is a “3D” visualization. As with most 3D visualizations, it is really 2D visualization with some perspective lines drawn on it to trick the brain. However, the user can interact with the visualization, clicking and dragging to *rotate* the visualization.



The visualization below has been rotated to best show the information in the 3rd dimension. It reveals that the main bit of information contained in the third dimension is None of these and the attributes of Sleepy and Innocent, and the two diet brands are closer to Feminine (keeping in mind that by "closer" we are referring to the angle to the origin).



Although 3D visualizations can be fun toys to show clients, a few points to keep in mind are:

- As mentioned above, 3D visualizations are really 2D.
- The optional 2D visualization is the default one produced by correspondence analysis. In the example shown in this chapter, while the "3D" visualization is showing us something that was not evident in the 2D variant, the cost of this is that it is showing less overall and is much less clear regarding the information in the first two dimensions (to appreciate this, contrast it with the traditional map at the beginning of the chapter).
- When using 3D visualizations, the subjectivity involved in the user rotating the map means that different people will end up drawing different conclusions. This is problematic, especially given the anecdotal evidence regarding the difficulty of many users to accurately interpret 2D correspondence analysis visualizations.

Software

To create an interactive 3D plot:

- Create a correspondence analysis in the usual way (see the earlier chapters).
- Either:
    - In Q: **Create > R Output**
    - In Displayr: **Insert > R Output**
- Paste in the following code, but replacing my.ca in the first two lines with the name of the correspondence analysis. (In Q, you find this by right-clicking on the correspondence analysis in the report tree and selecting **Reference name**).
- Press **Calculate**.

```
rc = my.ca$row.coordinates
cc = my.ca$column.coordinates
library(plotly)
p = plot_ly()
p = add_trace(p, x = rc[,1], y = rc[,2], z = rc[,3],
              mode = 'text', text = rownames(rc),
              textfont = list(color = "red"), showlegend = FALSE)
p = add_trace(p, x = cc[,1], y = cc[,2], z = cc[,3],
              mode = "text", text = rownames(cc),
              textfont = list(color = "blue"), showlegend = FALSE)
p <- config(p, displayModeBar = FALSE)
p <- layout(p, scene = list(xaxis = list(title = colnames(rc)[1]),
           yaxis = list(title = colnames(rc)[2]),
           zaxis = list(title = colnames(rc)[3]),
           aspectmode = "data"),
           margin = list(l = 0, r = 0, b = 0, t = 0))
p$sizingPolicy$browser$padding <- 0
my.3d.plot = p
```

Advanced topics

A basic problem with correspondence analysis is that some proportion of the variance in the data is not displayed in the visualizations, and this may cause the visualizations to be misleading.

Four approaches to addressing this issue are:

- Computing the quality of the map for each point (see **Quality**).
- **3D visualizations** (see the previous chapter).
- Supplementary data points
- Rotation

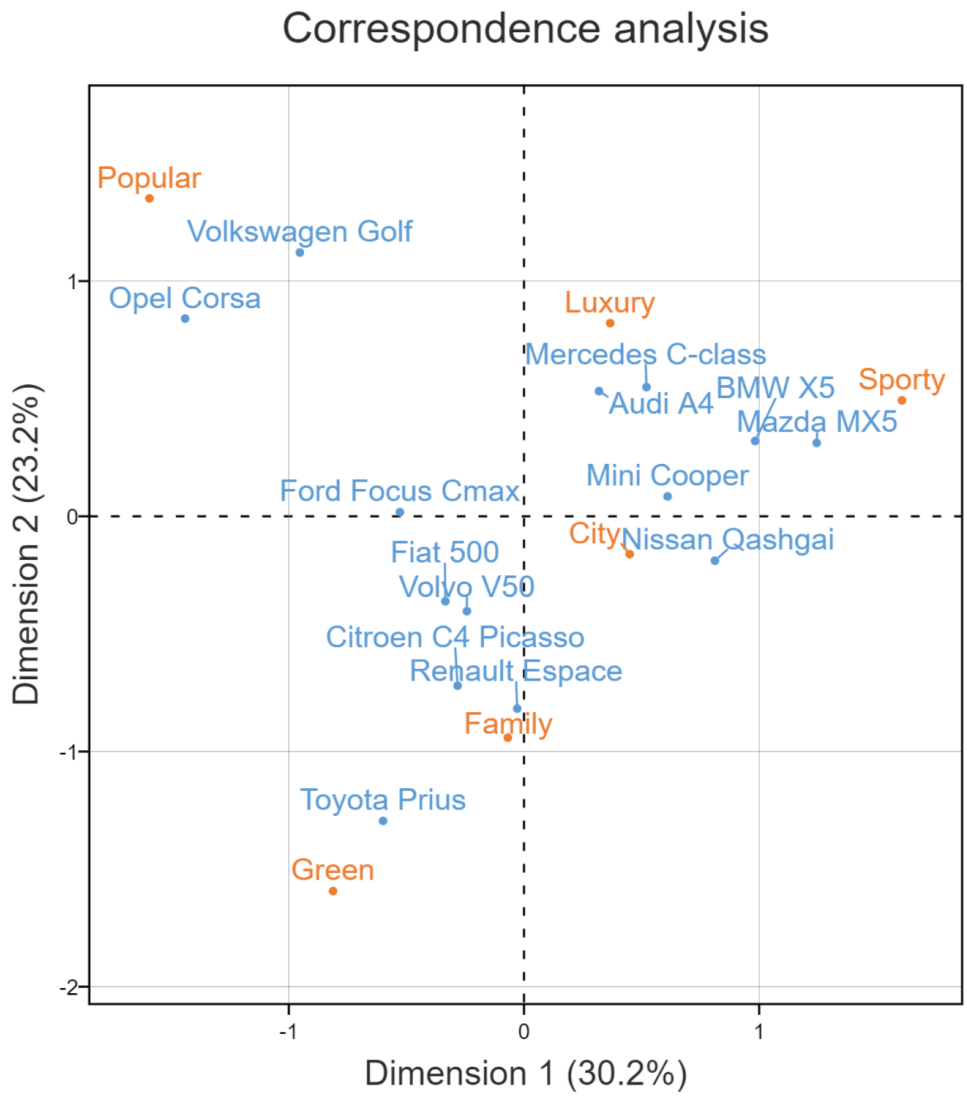
The last two of these are the focus of this chapter.

Supplementary data points

The use of supplementary data points refers to:

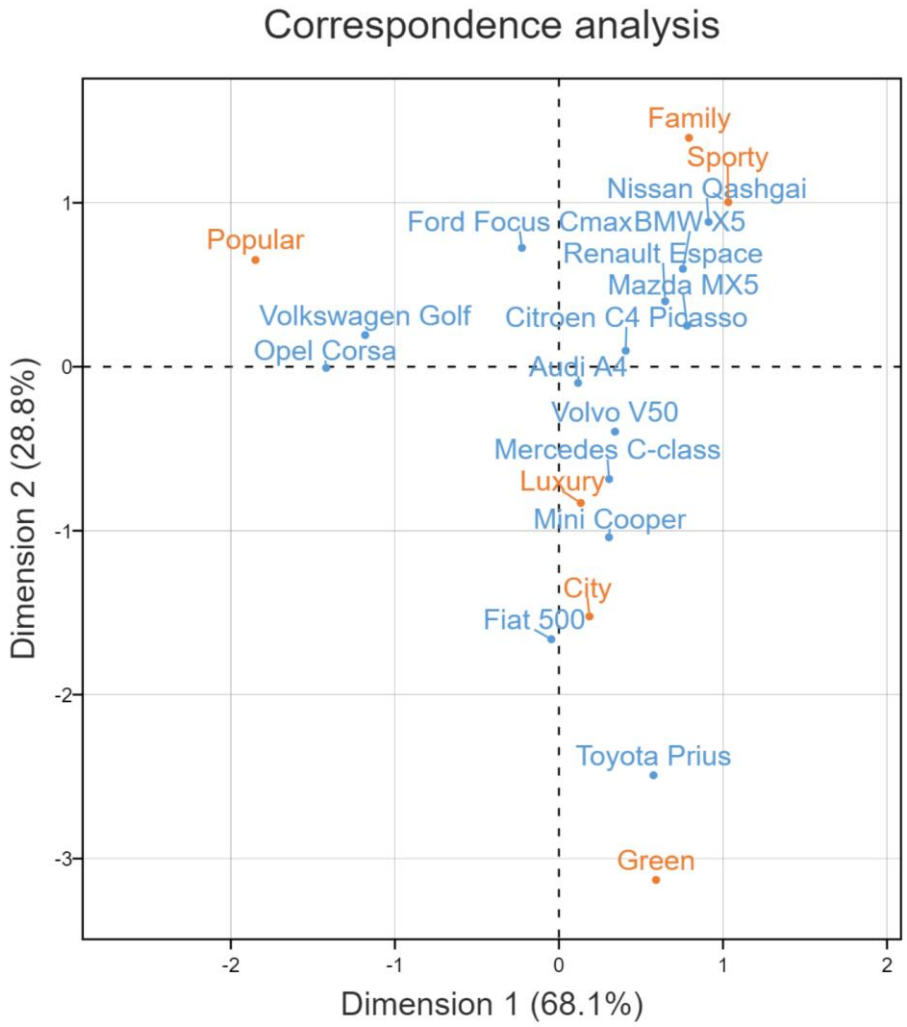
- Creating a correspondence analysis with only a subset of the data (the core data).
- Overlaying the unused data (the supplementary data) on the analysis created with the core data

The visualization below shows perceptions of 14 models of car. Let's say we wanted to study the four German brands. They form a line across the top from Volkswagen on the left, through Audi, Mercedes then BMW. We might be tempted to say that Volkswagen is Popular, Audi and Mercedes are Luxury and the BMW X5 is Sporty. However, the total explained variance is only 53%. This means there is information hidden in the dimensions that are not plotted.



In the correspondence analysis shown below, only the data for the four German brands has been used in working out the positions of these models and the attributes. The resulting map explains 97% of the variance in the data (the less data, the more variance explained, all else being equal). So, this visualization is much more accurate in terms of describing the relationship between these four models and the attributes. Now we see that the Audi A4 is very near the center of the plot. This means that it is not strongly associated with any of the characteristics when compared to the other German cars

(i.e., in the earlier visualization, it was luxurious when compared to all the brands, but is not when compared to the other German models).



The remaining brands have been overlain as *supplementary points,* allowing us to see their position in the same space that optimally describes the German models. In this case the analysis is ultimately not particularly successful, failing to pass the "smell test", with Family and Sporty positioned in the same place, and Green and Luxury also in the same direction. The reason that this analysis is not so wonderful is that ultimately the perceptional space that describes the positioning of the German models is different from that of the market as a whole, so the use of supplementary points means that we are no longer comparing like with like. The technique described in the next chapter, *rotation,* can be a better approach when there is a desire to focus on only a single row or column (e.g., brand) of a table.

Software

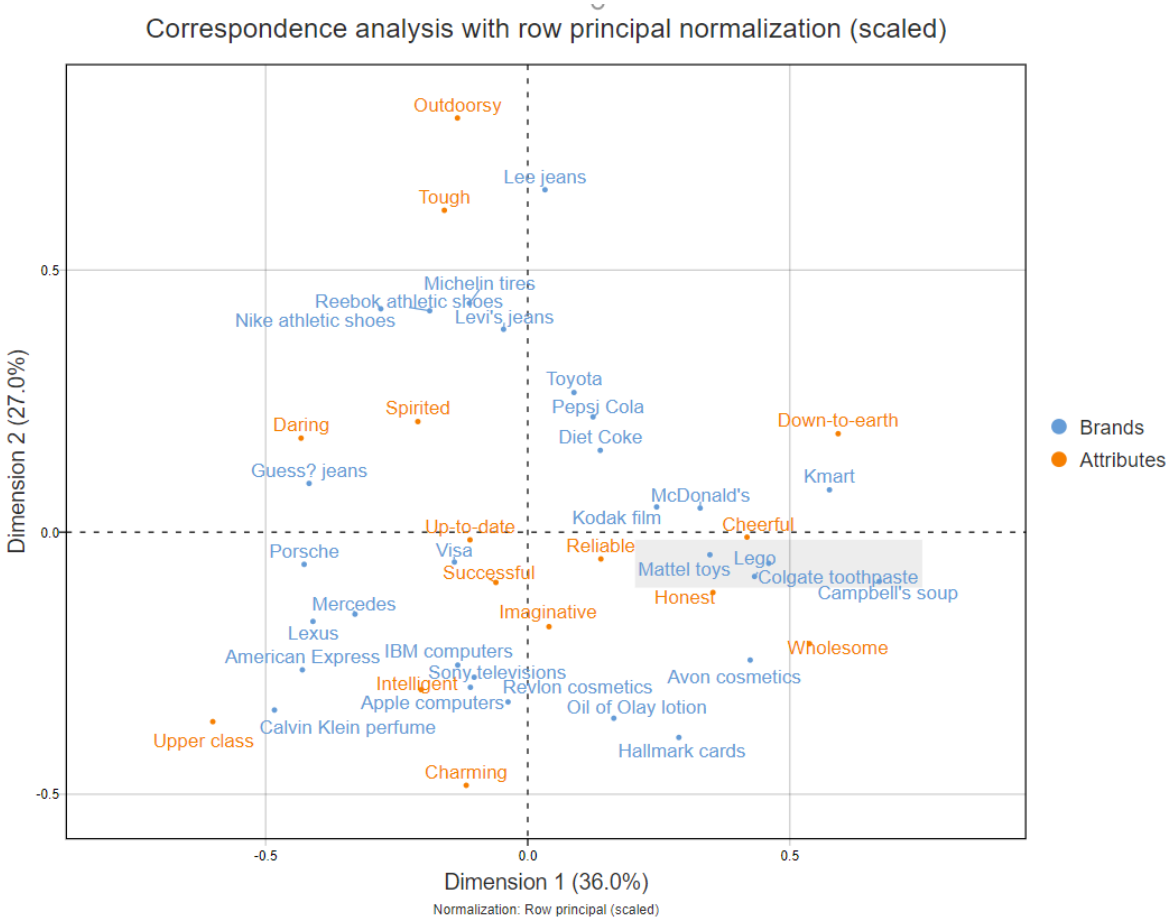
In Q and Displayr, the process for setting some of the points as supplementary is:

- Create the correspondence analysis in the usual way (see the earlier chapters)
- Type the names of the labels to be used as supplementary points, with commas separating them, in the **Supplementary** box (e.g., `Coke, Pepsi`).

Rotation

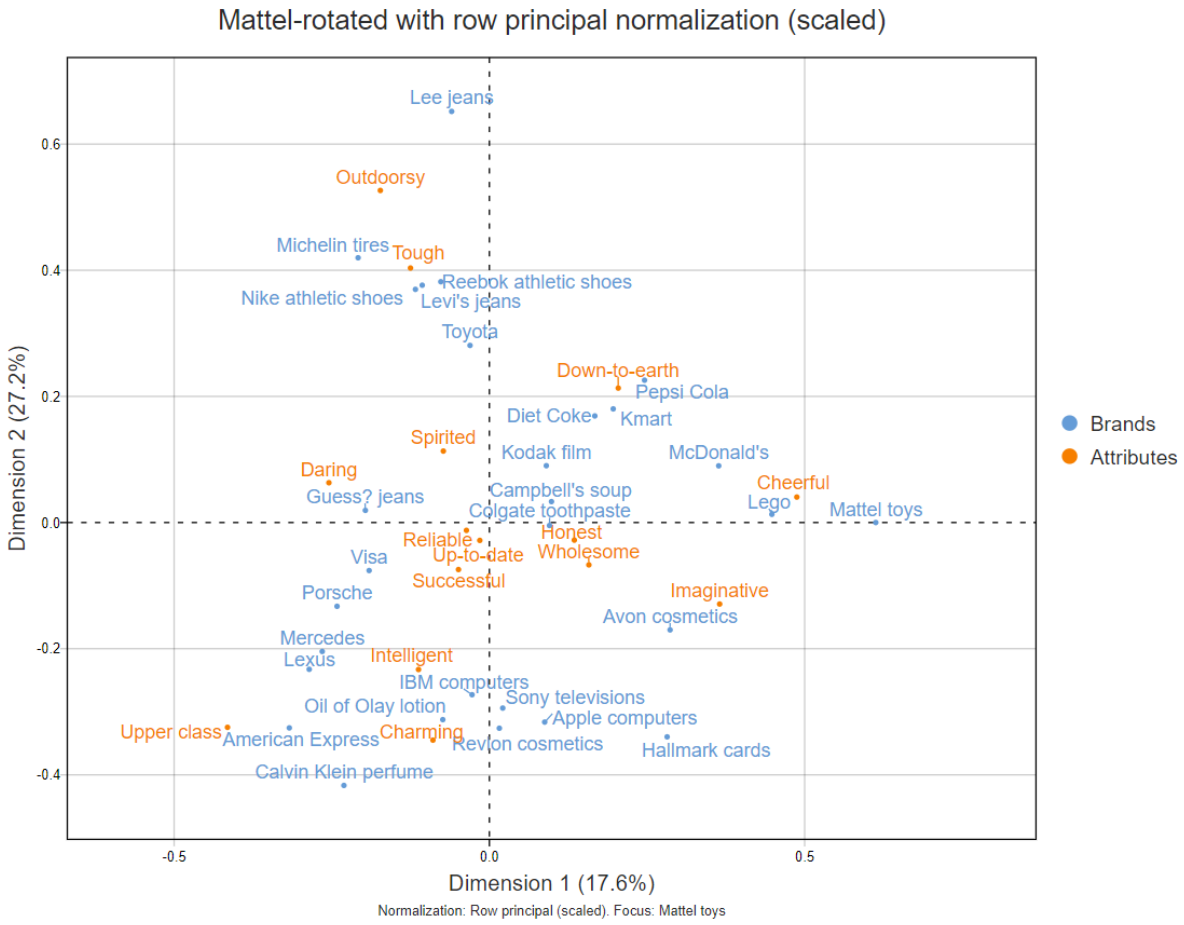
> *Rotation* is a technique that can be used to make a correspondence analysis optimally show the information for a single row or column of a table (e.g., a brand).

The plot below shows 28 brands by 15 attributes. The data table is shown in the first chapter of this eBook. Only the first two of 14 possible dimensions are plotted. These two dimensions explain 63% of the variance in the data, so inevitably there is some misrepresentation in the data.



If you look at the right-side of the map (around 3 O'clock from the center of the map), you can see that Mattel toys is adjacent to Colgate. This seems odd. It is possible to compute the variance explained for each of the brands (see **Quality**). In the case of Colgate toothpaste, the map explains 73% of the variance, which highlights that this map likely misrepresents the position of Colgate toothpaste in some ways. Only 33% of Mattel toys variance is explained by the map, suggesting its position may be

highly misleading. *Rotation* can be used to create a map that better represents the position of Mattel.[8] The map below has been created so that it explains 100% of the variance for Mattel. Mattel is now much further from the origin, as we would expect, and it is no longer so close to Colgate, which also makes sense. The resulting map is much better at displaying the position of Mattel in the market. However, the rotation is not without cost. It now only explains 45% of the variance in all the data, so while being better for Mattel, it is worse for most, or perhaps all, of the other brands (as always, the further the other brands from the origin, the more informative the map).



[8] Jake Hoare and Time Bock (Forthcoming), "A Brand's Eye View of Correspondence Analysis", International Journal of Research in Marketing.

Software

In Q and Displayr, the process for rotating the correspondence analysis to a specific row or column point is:

- Create the correspondence analysis in the usual way (see the earlier chapters)
- Type the names of the row or column label of interest in the **Focus** box