D – Question answered

Une image contenant table

Description générée automatiquement

First we try to take the Pearson's Chi-squared test. The output is : “X-squared = 570.04, df = 28, p-value < 2.2e-16”. The low p values tends to show that the data is corelated, so we can use CA to reduce the dimension.

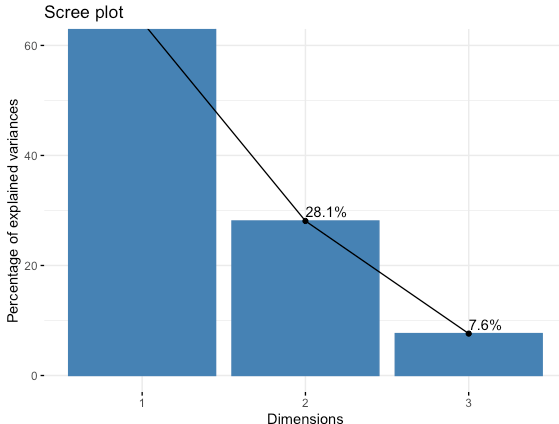
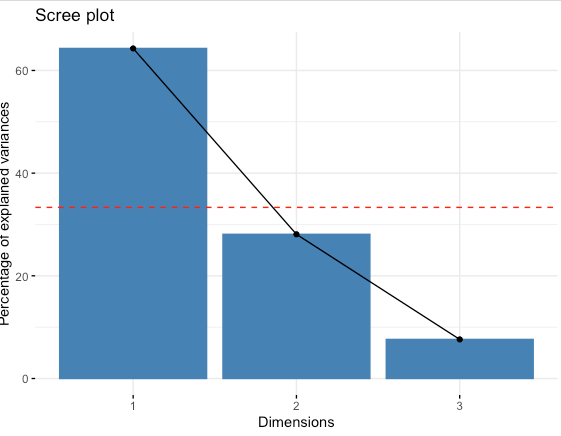
X-squared = 330.05, df = 21, p-value < 2.2e-16

We get the eigenvalues of the CA

Une image contenant texte

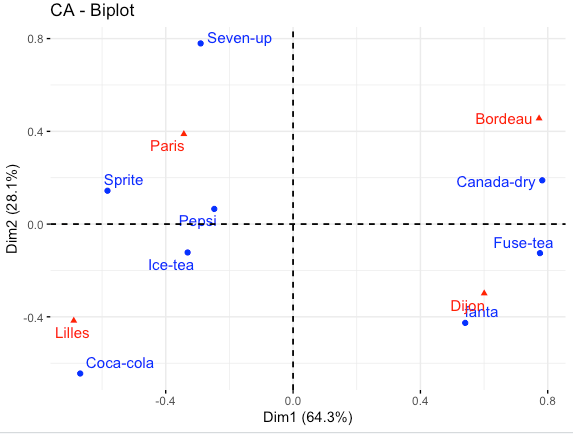
Description générée automatiquement

This output shows that 64% of the variance is described with the 1st dimension, and 28% with the second one for a total of 92%. Bellow, we can see a visualization of the eigenvalues. The other graph shows that each eigenvalue is supposed to represent 33% of the variance (we have three dimensions), and we can visualize that we can use the first two to show most of the variance in the data.

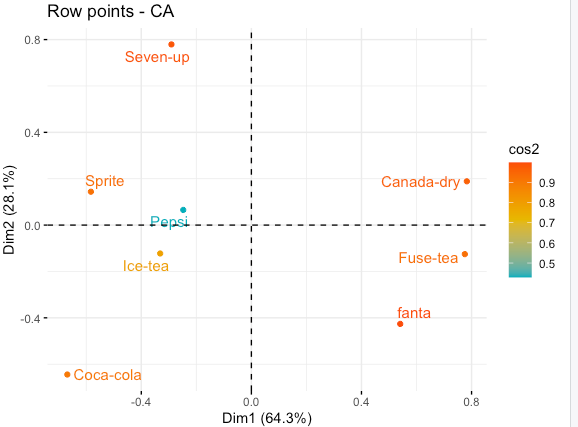


Then we create a biplot showing both rows and columns on a 2D graph. It is impossible to measure directly the distance between row or columns, we will have to use an asymmetric biplot to get information and interpret the distance between row and columns. However, we can interpret the distance between columns and columns, rows and rows which shows the similarity. Still the different areas give us information. We can see that Fanta and Fuse-tea are the most sell in Dijon, whereas other sodas like 7-UP are not selling there.

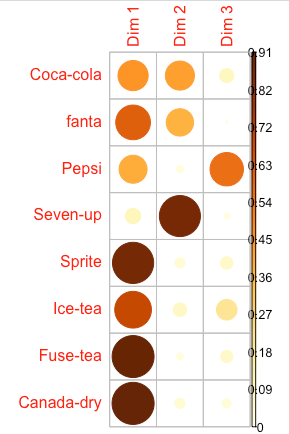
We can see that Paris and Dijon are in the opposite quartile, since they are not selling a lot of beverages in common.



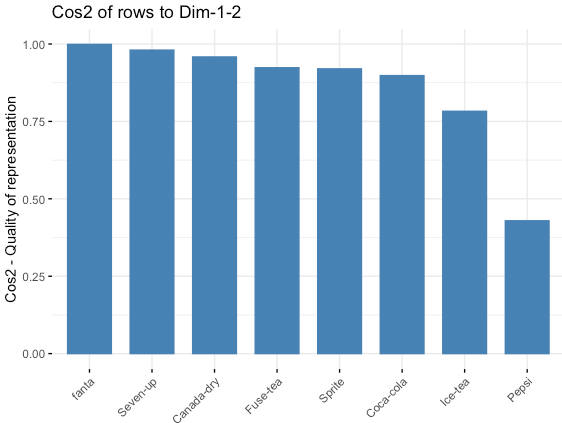
Here we can see the same plot (with only the rows) showing the cos2 value, which denotes how accurately the variable is shown. Here we can see that Fanta or seven UP are well describe over two dimension, but Pepsi isn’t.



Looking at this graph which shows the cos2 value of a column depending on the dimension we understand why: the Pepsi row is more expressed in the 3rd dimension than the 1st and 2nd, which is why the description on the two dimension plot is not optimal.



On the bar plot of cos2 we can see again that Pepsi should be interpreted with some caution. Indeed, we are lacking a dimension to express it correctly.

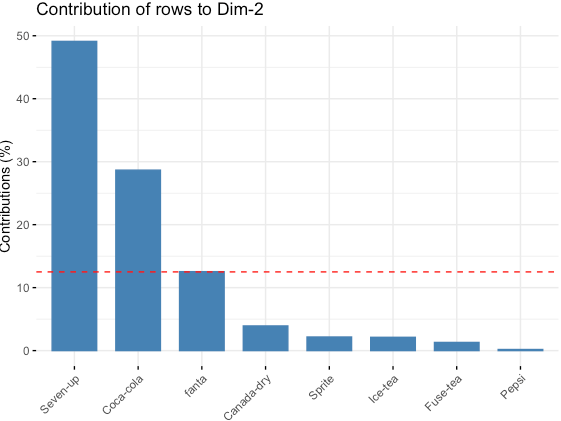


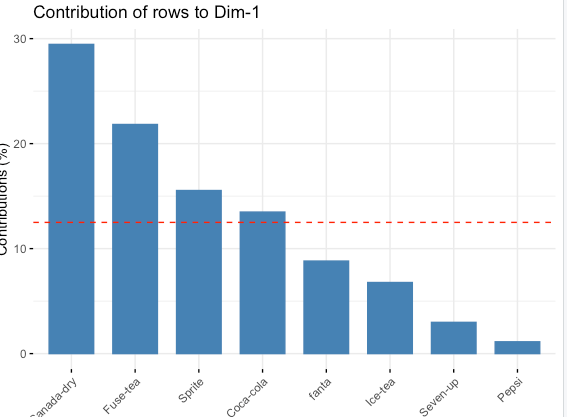
Below is the contribution of each from in the definition of the dimension. The rows that participate a lot in the first two dimension are the most important for our two-dimension description.

Une image contenant texte

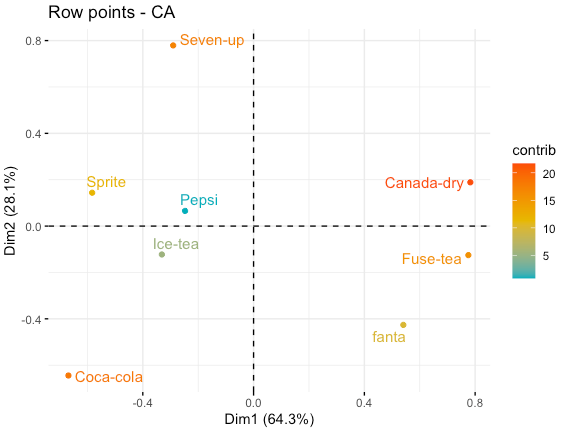
Description générée automatiquement

We can see that Canada-dry and Fuse-tea contribute the most to dim-1, and seven up and coca-cola to dim-2.

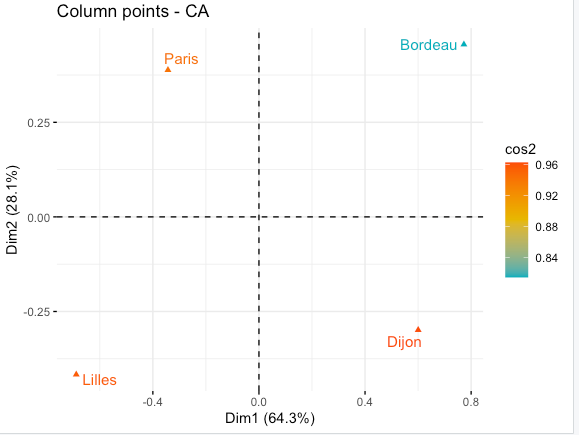


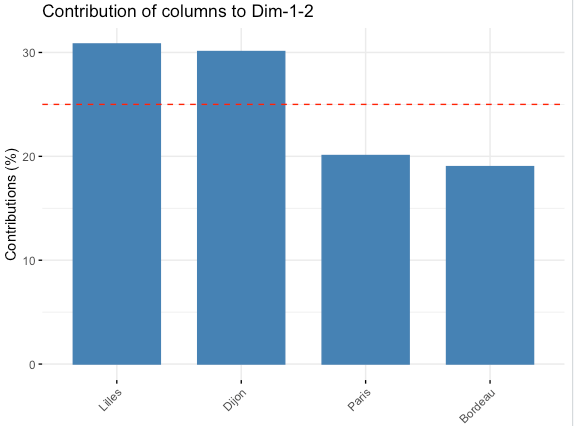


Bellow we can see that Seven up and Sprite have the most contribution on the negative side of dim-1, and Canada-Dry on the positive side which means dim-1 is defined by the opposition of those 3.



Now we will make the same work for columns. Bellow we can see the repartition of the different cities on the 2D plot, with their cos2 value. Again, we can see the Dijon and Lilles are very well represented but Bordeaux isn’t because its cos2 value is very low. This can also be highlighted on the second bar plot.





In the end, CA allowed us to reduce our dataset to two dimensions. We also understood why some values are better represented than others thanks to the cos2 values and the importance of some attributes in different dimensions. We also create a 2D plot allowing us to understand the link between different cities and the beverage they sell the most or didn’t really sell. We could get a better idea of how linked the cities and beverage sells are.