Principal Component Analysis

Dataset db_bloco_2_pos

This dataset contains 615 individuals and 7 variables.

---

1. Study of the outliers

The analysis of the graphs does not detect any outlier.

---

2. Inertia distribution

The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied.

The first two dimensions of analyse express **59.52%** of the total dataset inertia ; that means that 59.52% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is relatively high and thus the first plane well represents the data variability. This value is strongly greater than the reference value that equals **33.06%**, the variability explained by this plane is thus highly significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 4391 data tables of equivalent size on the basis of a normal distribution).

From these observations, it should be better to also interpret the dimensions greater or equal to the third one.

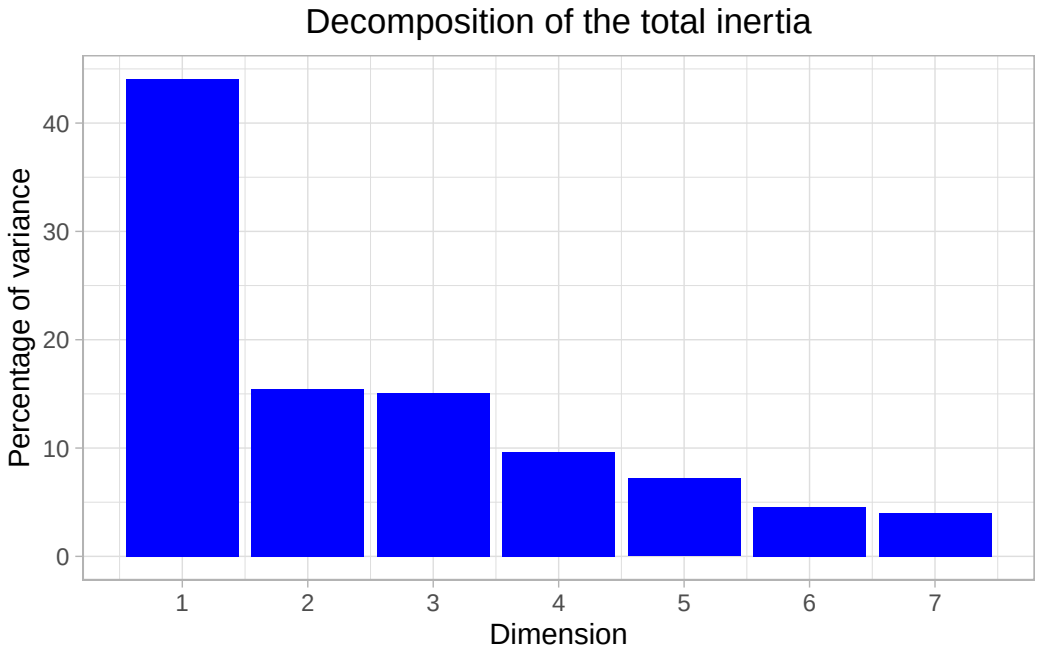

Figure 2 - Decomposition of the total inertia

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 1 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (44.07% against 17.27%). This observation suggests that only this axis is carrying a real information. As a consequence, the description will stand to these axis.

---

3. Description of the dimension 1

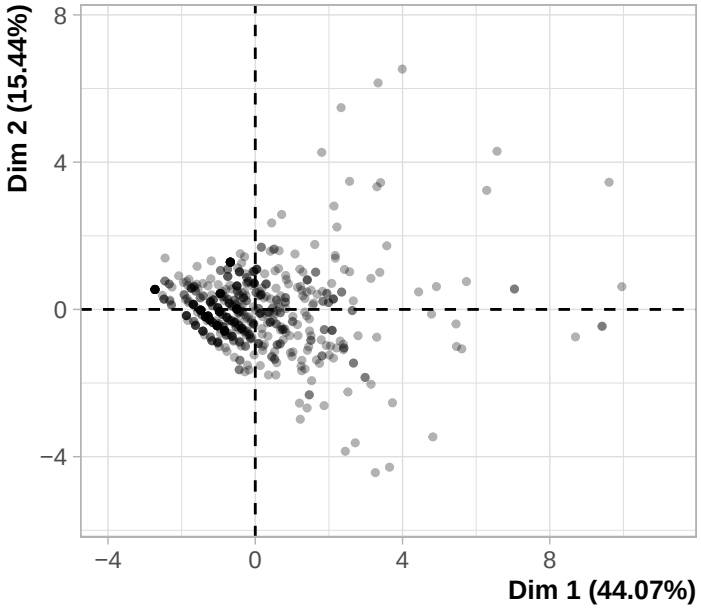

Figure 3.1 - Individuals factor map (PCA) *The labeled individuals are those with the higher contribution to the plane construction.*

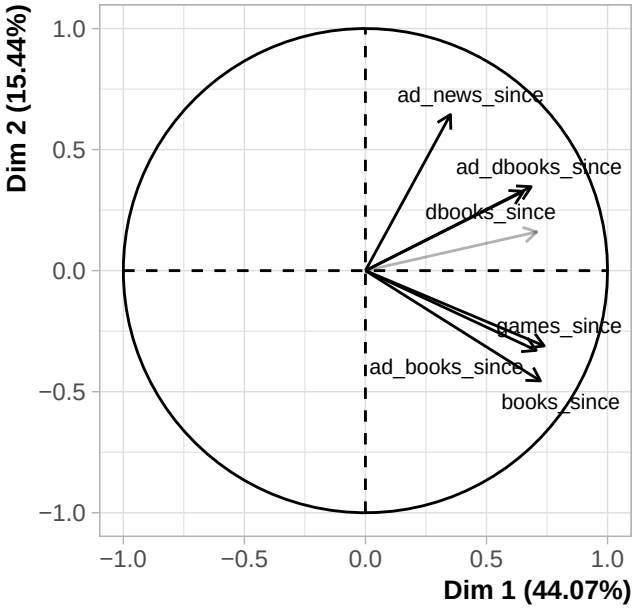

Figure 3.2 - Variables factor map (PCA) *The labeled variables are those the best shown on the plane.*

The **dimension 1** opposes individuals characterized by a strongly positive coordinate on the axis (to the right of the graph) to individuals characterized by a strongly negative coordinate on the axis (to the left of the graph).

The group 1 (characterized by a positive coordinate on the axis) is sharing :

- high values for the variables *ad_books_since*, *books_since*, *games_since*, *ad_dbooks_since*, *ad_news_since*, *dgames_since* and *dbooks_since* (variables are sorted from the strongest).

The group 2 (characterized by a positive coordinate on the axis) is sharing :

- high values for the variables *dbooks_since*, *dgames_since*, *ad_dbooks_since*, *games_since*, *books_since*, *ad_books_since* and *ad_news_since* (variables are sorted from the strongest).

The group 3 (characterized by a negative coordinate on the axis) is sharing :

- low values for the variables *ad_books_since*, *books_since*, *games_since*, *ad_dbooks_since*, *dgames_since*, *dbooks_since* and *ad_news_since* (variables are sorted from the weakest).

4. Classification

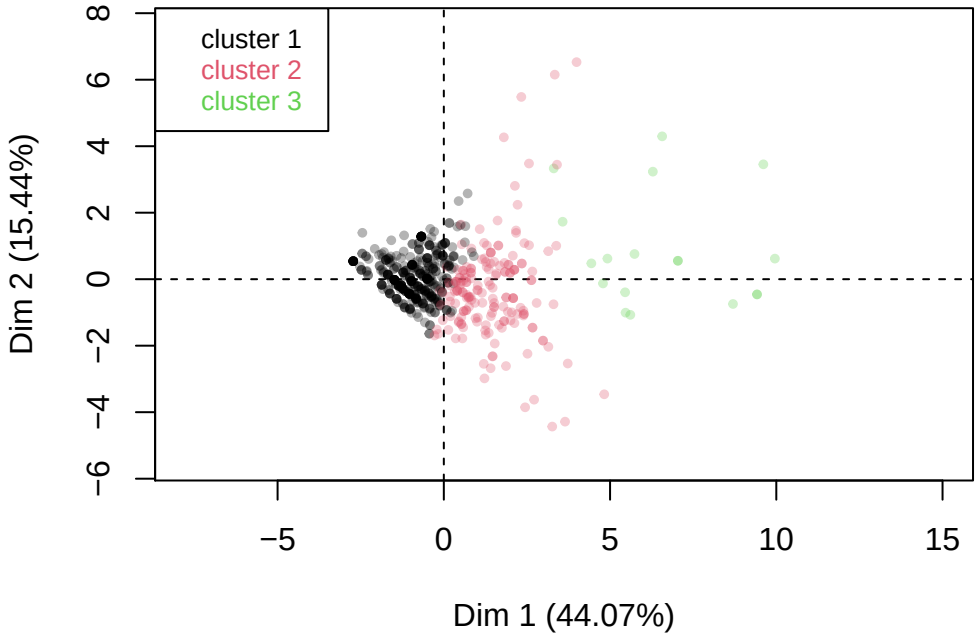

Figure 4 - Ascending Hierarchical Classification of the individuals. *The classification made on individuals reveals 3 clusters.*

The **cluster 1** is made of individuals sharing :

- low values for the variables *ad_books_since*, *books_since*, *ad_dbooks_since*, *games_since*, *dgames_since*, *dbooks_since* and *ad_news_since* (variables are sorted from the weakest).

The **cluster 2** is made of individuals sharing :

- high values for the variables *ad_books_since*, *books_since*, *ad_dbooks_since*, *games_since*, *ad_news_since*, *dgames_since* and *dbooks_since* (variables are sorted from the strongest).

The **cluster 3** is made of individuals sharing :

- high values for the variables *dbooks_since*, *dgames_since*, *ad_dbooks_since*, *games_since*, *books_since*, *ad_books_since* and *ad_news_since* (variables are sorted from the strongest).

Annexes