



Correspondence Analysis in Categorical Data

Sebastián David Reyes Santamaría

17 de marzo de 2025

1. Principal Component Analysis

We begin with the principal component analysis for each dataset (men and women). To do this, we perform the decomposition of the covariance matrix S into its eigenvectors and eigenvalues:

```
eigen() decomposition
$values
[1] 43.212400 14.958409 10.637096 2.629595

$vectors
      [,1]      [,2]      [,3]
[1,] -0.2245473 -0.06732947 0.2398902
[2,] -0.3155909 0.06695456 0.8980264
[3,] -0.7106910 -0.63913767 -0.2523383
[4,] -0.5872811 0.76320827 -0.2689361
      [,4]
[1,] 0.94207110
[2,] -0.29911201
[3,] -0.15081989
[4,] -0.01695292
```

Figura 1: Decomposition for men

```
eigen() decomposition
$values
[1] 48.955644 18.463514 13.537986 4.818058

$vectors
      [,1]      [,2]      [,3]
[1,] -0.2173746 -0.2730167 0.3733308
[2,] -0.3878472 -0.6205057 0.4658646
[3,] -0.6807847 -0.1707259 -0.7076914
[4,] -0.5821125 0.7150435 0.3778455
      [,4]
[1,] 0.85955471
[2,] -0.49751123
[3,] 0.08097994
[4,] -0.08420525
```

Figura 2: Decomposition for women

The eigenvalues represent the amount of variance explained by each principal component. In this case, the values are higher for the women's dataset, suggesting that the variables might be more correlated.

Visualizing the PCA analysis along with the observations of each dataset, we obtain:

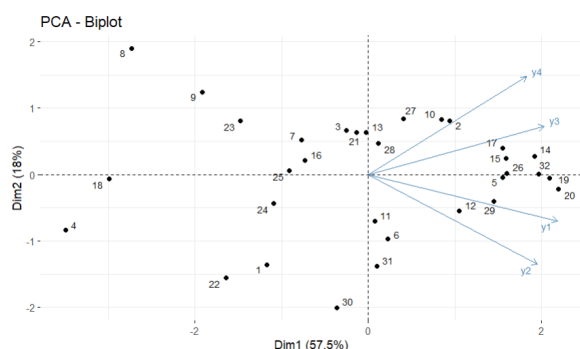


Figura 3: PCA for men

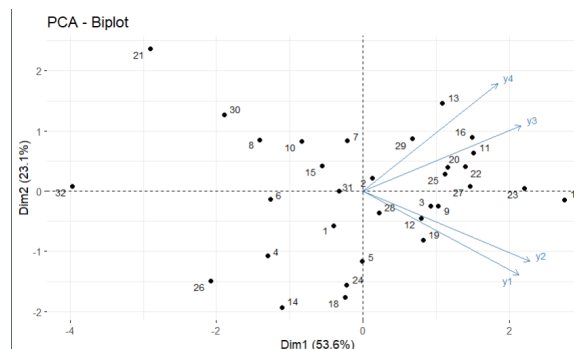


Figura 4: PCA for women

Both analyses capture approximately 75 % of the total variability of the data in their first two dimensions. According to the results, the variables seem to form two main groups: (y_3, y_4) and (y_1, y_2) . In the first component, the explained variance is higher for men (57.50 %) than for women (53.56 %), whereas in the second component, the explained variance is higher for women (23.10 %) than for men (17.97 %).

Selection of the number of principal components

To determine the number of principal components to retain, various criteria were used, such as the rule of eigenvalues greater than 1, the elbow method, and the parallel method. As a result, it was decided to retain only one principal component in both cases.

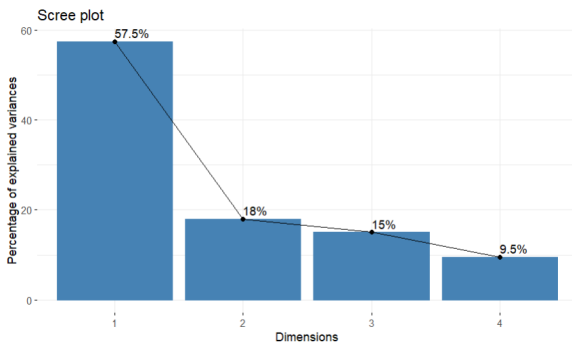


Figura 5: Elbow method for men

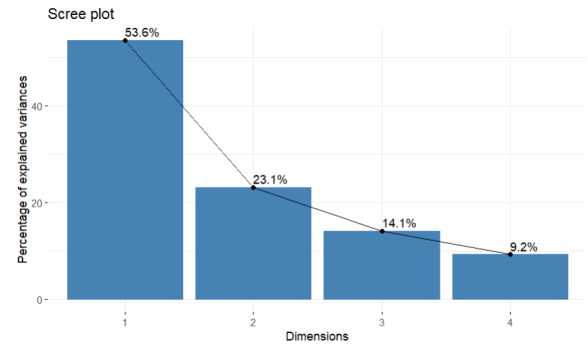


Figura 6: Elbow method for women

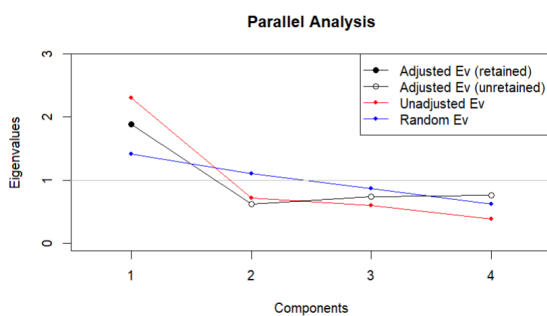


Figura 7: Parallel method for men

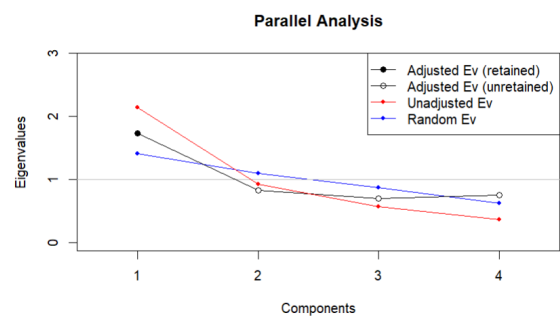


Figura 8: Parallel method for women

Under the first criterion of considering components with eigenvalues greater than 1, there is some doubt about the eigenvalue of the second component in the PCA for women, which takes a value of 0,924 (the closest to 1). However, after analyzing with the other two criteria, it is still suggested to select only one component. With the elbow method, a clear break is observed for men in the first component, while for women, it is less pronounced but still follows the expected pattern. Also, Horn's parallel analysis suggests retaining only one component in both cases, as the random eigenvalue line crosses the other lines after the first component.

Interpretation of the Principal Component

The first component is characterized by a clear opposition between two groups of variables:

- y_3 and y_4 are positively correlated with each other and have a positive correlation with Dim1.
- y_1 and y_2 are positively correlated with each other and have a negative correlation with Dim1.



Individuals on the positive side of Dim1 (right) tend to have high values in y3 and y4, and low values in y1 and y2. Conversely, individuals on the negative side of Dim1 (left) tend to have high values in y1 and y2, and low values in y3 and y4.

Contribution of variables to the principal component

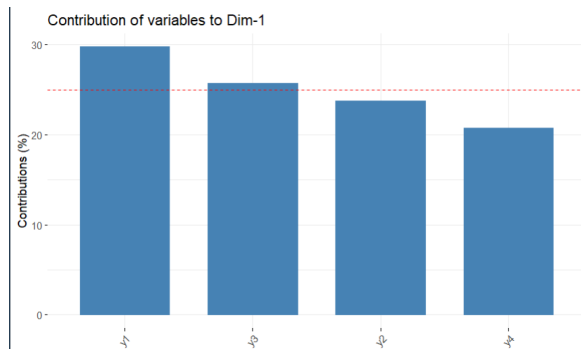


Figura 9: Contribution of each variable in the first principal component for men

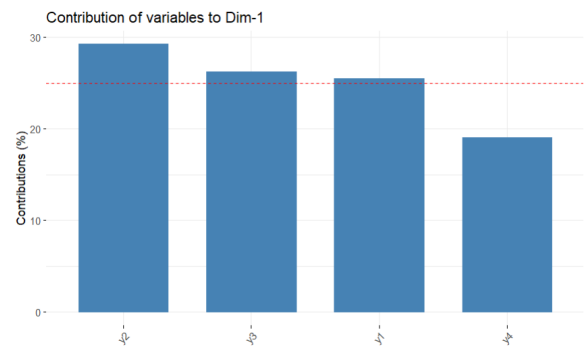


Figura 10: Contribution of each variable in the first principal component for women

The dominant variable changes: y1 in the first dataset vs. y2 in the second. y4 consistently shows the lowest contribution, although it remains relevant, suggesting that the first principal component captures a robust pattern in the data.

2. Correspondence Analysis

For correspondence analysis, contingency tables are first entered into R:

Clothing	Clothing accessories	Footwear	Accessories	Books	Records	Household goods	Seasons	Types	Seasons	Perfumes	Hidden books	Other	Total
112	19	35	224	19	7	22	127	119	142	79	15	24	642
113	41	48	111	249	60	37	21	140	19	149	19	19	208
114	471	114	38	91	50	27	41	88	14	363	141	9	1471
115	456	108	76	18	32	12	33	52	19	74	79	14	127
116	100	100	100	100	100	100	100	100	100	100	100	100	1000
117	100	100	100	100	100	100	100	100	100	100	100	100	1000
118	100	100	100	100	100	100	100	100	100	100	100	100	1000
119	100	100	100	100	100	100	100	100	100	100	100	100	1000
120	100	100	100	100	100	100	100	100	100	100	100	100	1000
121	100	100	100	100	100	100	100	100	100	100	100	100	1000
122	100	100	100	100	100	100	100	100	100	100	100	100	1000
123	100	100	100	100	100	100	100	100	100	100	100	100	1000
124	100	100	100	100	100	100	100	100	100	100	100	100	1000
125	100	100	100	100	100	100	100	100	100	100	100	100	1000
126	100	100	100	100	100	100	100	100	100	100	100	100	1000
127	100	100	100	100	100	100	100	100	100	100	100	100	1000
128	100	100	100	100	100	100	100	100	100	100	100	100	1000
129	100	100	100	100	100	100	100	100	100	100	100	100	1000
130	100	100	100	100	100	100	100	100	100	100	100	100	1000
131	100	100	100	100	100	100	100	100	100	100	100	100	1000
132	100	100	100	100	100	100	100	100	100	100	100	100	1000
133	100	100	100	100	100	100	100	100	100	100	100	100	1000
134	100	100	100	100	100	100	100	100	100	100	100	100	1000
135	100	100	100	100	100	100	100	100	100	100	100	100	1000
136	100	100	100	100	100	100	100	100	100	100	100	100	1000
137	100	100	100	100	100	100	100	100	100	100	100	100	1000
138	100	100	100	100	100	100	100	100	100	100	100	100	1000
139	100	100	100	100	100	100	100	100	100	100	100	100	1000
140	100	100	100	100	100	100	100	100	100	100	100	100	1000
141	100	100	100	100	100	100	100	100	100	100	100	100	1000
142	100	100	100	100	100	100	100	100	100	100	100	100	1000
143	100	100	100	100	100	100	100	100	100	100	100	100	1000
144	100	100	100	100	100	100	100	100	100	100	100	100	1000
145	100	100	100	100	100	100	100	100	100	100	100	100	1000
146	100	100	100	100	100	100	100	100	100	100	100	100	1000
147	100	100	100	100	100	100	100	100	100	100	100	100	1000
148	100	100	100	100	100	100	100	100	100	100	100	100	1000
149	100	100	100	100	100	100	100	100	100	100	100	100	1000
150	100	100	100	100	100	100	100	100	100	100	100	100	1000
151	100	100	100	100	100	100	100	100	100	100	100	100	1000
152	100	100	100	100	100	100	100	100	100	100	100	100	1000
153	100	100	100	100	100	100	100	100	100	100	100	100	1000
154	100	100	100	100	100	100	100	100	100	100	100	100	1000
155	100	100	100	100	100	100	100	100	100	100	100	100	1000
156	100	100	100	100	100	100	100	100	100	100	100	100	1000
157	100	100	100	100	100	100	100	100	100	100	100	100	1000
158	100	100	100	100	100	100	100	100	100	100	100	100	1000
159	100	100	100	100	100	100	100	100	100	100	100	100	1000
160	100	100	100	100	100	100	100	100	100	100	100	100	1000
161	100	100	100	100	100	100	100	100	100	100	100	100	1000
162	100	100	100	100	100	100	100	100	100	100	100	100	1000
163	100	100	100	100	100	100	100	100	100	100	100	100	1000
164	100	100	100	100	100	100	100	100	100	100	100	100	1000
165	100	100	100	100	100	100	100	100	100	100	100	100	1000
166	100	100	100	100	100	100	100	100	100	100	100	100	1000
167	100	100	100	100	100	100	100	100	100	100	100	100	1000
168	100	100	100	100	100	100	100	100	100	100	100	100	1000
169	100	100	100	100	100	100	100	100	100	100	100	100	1000
170	100	100	100	100	100	100	100	100	100	100	100	100	1000
171	100	100	100	100	100	100	100	100	100	100	100	100	1000
172	100	100	100	100	100	100	100	100	100	100	100	100	1000
173	100	100	100	100	100	100	100	100	100	100	100	100	1000
174	100	100	100	100	100	100	100	100	100	100	100	100	1000
175	100	100	100	100	100	100	100	100	100	100	100	100	1000
176	100	100	100	100	100	100	100	100	100	100	100	100	1000
177	100	100	100	100	100	100	100	100	100	100	100	100	1000
178	100	100	100	100	100	100	100	100	100	100	100	100	1000
179	100	100	100	100	100	100	100	100	100	100	100	100	1000
180	100	100	100	100	100	100	100	100	100	100	100	100	1000
181	100	100	100	100	100	100	100	100	100	100	100	100	1000
182	100	100	100	100	100	100	100	100	100	100	100	100	1000
183	100	100	100	100	100	100	100	100	100	100	100	100	1000
184	100	100	100	100	100	100	100	100	100	100	100	100	1000
185	100	100	100	100	100	100	100	100	100	100	100	100	1000
186	100	100	100	100	100	100	100	100	100	100	100	100	1000
187	100	100	100	100	100	100	100	100	100	100	100	100	1000
188	100	100	100	100	100	100	100	100	100	100	100	100	1000
189	100	100	100	100	100	100	100	100	100	100	100	100	1000
190	100	100	100	100	100	100	100	100	100	100	100	100	1000
191	100	100	100	100	100	100	100	100	100	100	100	100	1000
192	100	100	100	100	100	100	100	100	100	100	100	100	1000
193	100	100	100	100	100	100	100	100	100	100	100	100	1000
194	100	100	100	100	100	100	100	100	100	100	100	100	1000
195	100	100	100	100	100	100	100	100	100	100	100	100	1000
196	100	100	100	100	100	100	100	100	100	100	100	100	1000
197	100	100	100	100	100	100	100	100	100	100	100	100	1000
198	100	100	100	100	100	100	100	100	100	100	100	100	1000
199	100	100	100	100	100	100	100	100	100	100	100	100	1000
200	100	100	100	100	100	100	100	100	100	100	100	100	1000

Figura 11: Contingency table for womens

Clothing	Clothing accessories	Footwear, shoes	Accessories	Books	Records	Household goods	Seasons	Types	Seasons	Perfume	Hidden books	Other	Total
112	19	35	224	19	7	22	127	119	142	79	15	24	642
113	41	48	111	249	60	37	21	140	19	149	19	19	208
114	471	114	38	91	50	27	41	88	14	363	141	9	1471
115	456	108	76	18	32	12	3						
116	134	138	204	380	149	25	272	217	637	664	48	37	347
117	304	193	259	527	258	368	89	246	116	296	61	492	454
118	320	304	143	551	84	165	61	40	15	71	52	138	252
119	342	342	342	342	342	342	342	342	342	342	342	342	342
120	350	350	350	350	350	350	350	350	350	350	350	350	350
121	350	350	350	350	350	350	350	350	350	350	350	350	350
122	350	350	350	350	350	350	350	350	350	350	350	350	350
123	350	350	350	350	350	350	350	350	350	350	350	350	350
124	350	350	350	350	350	350	350	350	350	350	350	350	350
125	350	350	350	350	350	350	350	350	350	350	350	350	350
126	350	350	350	350	350	350	350	350	350	350	350	350	350
127	350	350	350	350	350	350	350	350	350	350	350	350	350
128	350	350	350	350	350	350	350	350	350	350	350	350	350
129	350	350	350	350	350	350	350	350	350	350	350	350	350
130	350	350	350	350	350	350	350	350	350	350	350	350	350
131	350	350	350	350	350	350	350	350	350	350	350	350	350
132	350	350	350	350	350	350	350	350	350	350	350	350	350
133	350	350	350	350	350	350	350	350	350	350	350	350	350
134	350	350	350	350	350	350	350	350	350	350	350	350	350
135	350	350	350	350	350	350	350	350	350	350	350	350	350
136	350	350	350	350	350	350	350	350	350	350	350	350	350
137	350	350	350	350	350	350	350	350	350	350	350	350	350
138	350	350	350	350	350	350	350	350	350	350	350	350	350
139	350	350	350	350	350	350	350	350	350	350	350	350	350
140	350	350	350	350	350	350	350	350	350	350	350	350	350
141	350	350	350	350	350	350	350	350	350	350	350	350	350
142	350	350	350	350	350	350	350	350	350	350	350	350	350
143	350	350	350	350	350	350	350	350	350	350	350	350	350
144	350	350	350	350	350	350	350	350	350	350	350	350	350
145	350	350	350	350	350	350	350	350	350	350	350	350	350
146	350	350	350	350	350	350	350	350	350	350	350	350	350
147	350	350	350	350	350	350	350	350	350	350	350	350	350
148	350	350	350	350	350	350	350	350	350	350	350	350	350
149	350	350	350	350	350	350	350	350	350	350	350	350	350
150	350	350	350	350	350	350	350	350	350	350	350	350	350
151	350	350	350	350	350	350	350	350	350	350	350	350	350
152	350	350	350	350	350	350	350	350	350	350	350	350	350
153	350	350	350	350	350	350	350	350	350	350	350	350	350
154	350	350	350	350	350	350	350	350	350	350	350	350	350
155	350	350	350	350	350	350	350	350	350	350	350	350	350
156	350	350	350	350	350	350	350	350	350	350	350	350	350
157	350	350	350	350	350	350	350	350	350	350	350	350	350
158	350	350	350	350	350	350	350	350	350	350	350	350	350
159	350	350	350	350	350	350	350	350	350	350	350	350	350
160	350	350	350	350	350	350	350	350	350	350	350	350	350
161	350	350	350	350	350	350	350	350	350	350	350	350	350
162	350	350	350	350	350	350	350	350	350	350	350	350	350
163	350	350	350	350	350	350	350	350	350	350	350	350	350
164	350	350	350	350	350	350	350	350	350	350	350	350	350
165	350	350	350	350	350	350	350	350	350	350	350	350	350
166	350	350	350	350	350	350	350	350	350	350	350	350	350
167	350	350	350	350	350	350	350	350	350	350	350	350	350
168	350	350	350	350	350	350	350	350	350	350	350	350	350
169	350	350	350	350	350	350	350	350	350	350	350	350	350
170	350	350	350	350	350	350	350	350	350	350	350	350	350
171	350	350	350	350	350	350	350	350	350	350	350	350	350
172	350	350	350	350	350	350	350	350	350	350	350	350	350
173	350	350	350	350	350	350	350	350	350	350	350	350	350
174	350	350	350	350	350	350	350	350	350	350	350	350	350
175	350	350	350	350	350	350	350	350	350	350	350	350	350
176	350	350	350	350	350	350	350	350	350	350	350	350	350
177	350	350	350	350	350	350	350	350	350	350	350	350	350
178	350	350	350	350	350	350	350	350	350	350	350	350	350
179	350	350	350	350	350	350	350	350	350	350	350	350	350
180	350	350	350	350	350	350	350	350	350	350	350	350	350
181	350	350	350	350	350	350	350	350	350	350	350	350	350
182	350	350	350	350	350	350	350	350	350	350	350	350	350
183	350	350	350	350	350	350	350	350	350	350	350	350	350
184	350	350	350	350	350	350	350	350	350	350	350	350	350
185	350	350	350	350	350	350	350	350	350	350	350	350	350
186	350	350	350	350	350	350	350	350	350	350	350	350	350
187	350	350	350	350	350	350	350	350	350	350	350	350	350
188	350	350	350	350	350	350	350	350	350	350	350	350	350
189	350	350	350	350	350	350	350	350	350	350	350	350	350
190	350	350	350	350	350	350	350	350	350	350	350	350	350
191	350	350	350	350	350	350	350	350	350	350	350	350	350
192	350	350	350	350	350	350	350	350	350	350	350	350	350
193	350	350	350	350	350	350	350	350	350	350	350	350	350
194	350	350	350	350	350	350	350	350	350	350	350	350	350
195	350	350	350	350	350	350	350	350	350	350	350	350	350
196	350	350	350	350	350	350	350	350	350	350	350	350	350
197	350	350	350	350	350	350	350	350	350	350	350	350	350
198	350	350	350	350	350	350	350	350	350	350	350	350	350
199	350	350	350	350	350	350	350	350	350	350	350	350	350
200	350	350	350	350	350	350	350	350	350	350	350	350	350
201	350	350	350	350	350	350	350	350	350	350	350	350	350
202	350	350	350	350	350	350	350	350	350	350	350	350	350
203	350	350	350	350	350	350	350	350	350	350	350	350	350
204	350	350	350	350	350	350	350	350	350	350	350	350	350
205	350	350	350	350	350	350	350	350	350	350	350	350	350
206	350	350	350	350	350	350	350	350	350	350	350	350	350
207	350	350	350	350	350	350	350	350	350	350	350	350	350
208	350	350	350	350	350	350	350	350	350	350	350	350	350
209	350	350	350	350	350	350	350	350	350	350	350	350	350
210	350	350	350	350	350	350	350	350	350	350	350	350	350
211	350	350	350	350	350	350	350	350	350	350	350	350	350
212	350	350	350	350	350	350	350	350	350	350	350	350	350
213	350	350	350	350	350	350	350	350	350	350	350	350	350
214	350	350	350	350	350	350	350	350	350	350	350	350	350
215	350	350	350	350	350	350	350	350	350	350	350	350	350
216	350	350	350	350	350	350	350	350	350	350	350	350	350
217	350	350	350	350	350	350	350	350	350	350	350	350	350
218	350	350	350	350	350								



data, without including dimensions that contribute very little explained variance, we will select these first 2 dimensions for the analysis.

Main findings

Starting the analysis, we see that the age categories (<12) and (12-14) have the highest contributions to Dimension 1, suggesting that these categories are strongly associated with the (Items) categories in this dimension. The ages (+65) and (50-64) have considerable contributions to Dimension 2, indicating a significant association with different (Items) in this dimension.

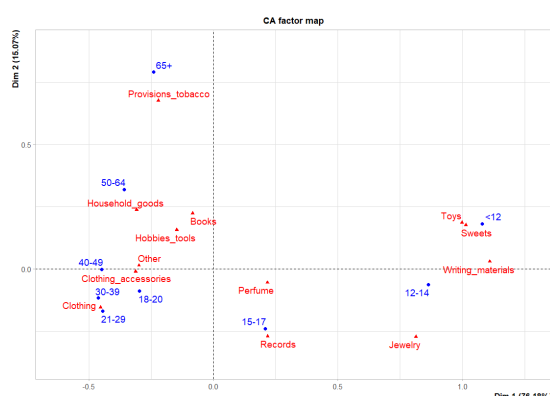


Figura 13: Correspondence analysis for women

When interpreting Dimension 1, it seems to be more related to age group differences, especially with the older age group (65+), suggesting that people aged 65 or older have distinctive characteristics or preferences compared to other age groups. The younger categories, (15-17) and (18-20), have lower contributions in this dimension. Dimension 2 is more influenced by the Item (Provisions tobacco), in the ages (65+) and (50-64) showing a higher association.

We can see relationships between the categories, for example, that the recorded thefts for ages (<12) and (12-14) tend to be from the Items (Toys, Sweets, Writing materials), for the age (15-17) the theft of the item (Records) is more common, ages (21-29) and (30-39) with the Item (Clothing), age (50-64) with (Household goods) and finally age (65+) with (Provisions tobacco).

2.2. Correspondence analysis for men

How many dimensions to consider:

Similar to the previous analysis, Dimension 1 explains 77.10 % of the variance, followed



by Dimension 2 with 11.45 %. Together, the first two dimensions explain 88.55 % of the total variance in the data. We will select these first 2 dimensions for the analysis.

Main findings

The age categories (<12, 12-14, 20-29) and the Item (Toys) have the highest contributions to Dimension 1. On the other hand, the age (15-17) and the Item (Records) have a considerable contribution to Dimension 2.

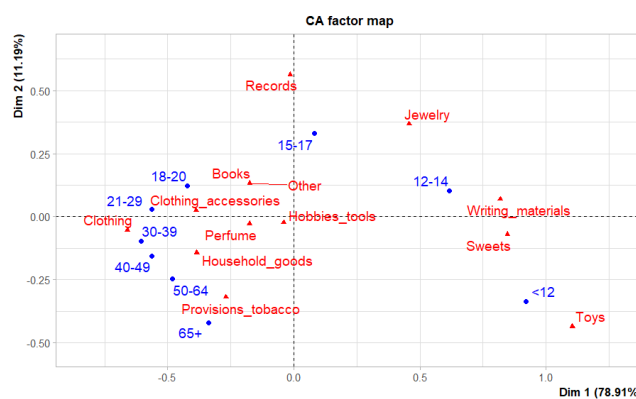


Figura 14: Correspondence analysis for men

The younger ages (<12, 12-14) appear to be more related to the Items (Toys, Sweets, Writing materials). On the other hand, the ages (21-29, 30-39) seem to be more associated with the Item (Clothing). Moreover, the older ages, particularly (50-64, 65+) show a strong association with the Items (Household goods and Provisions tobacco), similar to the case of women.

2.3. Joint correspondence analysis

Although the positions and contributions of the categories to each dimension vary, in general, a similar association between the different categories is maintained, as seen in the graph. Likewise, the first two dimensions explain 90.1 % of the variance in the data.

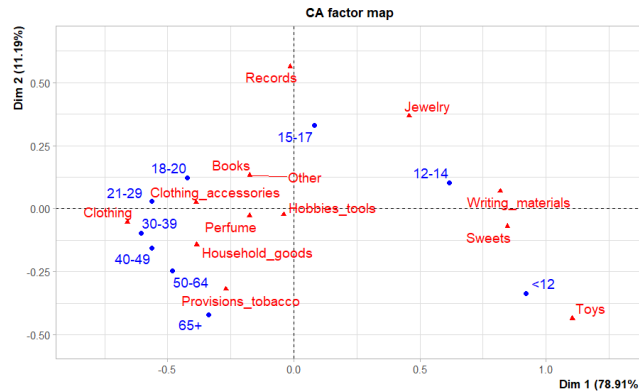


Figura 15: Joint correspondence analysis

The strongest associations remain, for example, for age (<12) being closer to the Item (Toys), age (12-14) to the Items (Sweets and Writing materials) or age (65+) to (Provisions tobacco), in general, maintaining the relationships given in the two separate correspondence analyses for men and women.

3. Code

Click here to go to the GitHub repository with the code

File name: CA_analysis_gender_age_groups.R

Referencias

Heijden, P. G. M., Falguerolles, A., y De Leeuw, J. (2018). A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Loglinear Analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 38(2), 249-273. <https://doi.org/10.2307/2348058>