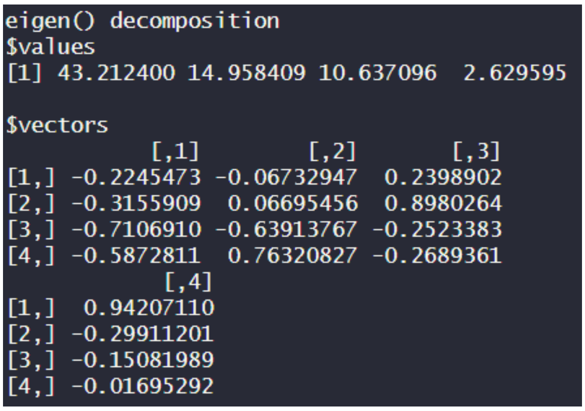
Correspondence Analysis in Categorical Data

Sebastián David Reyes Santamaría

March 19, 2025

1 Principal Component Analysis

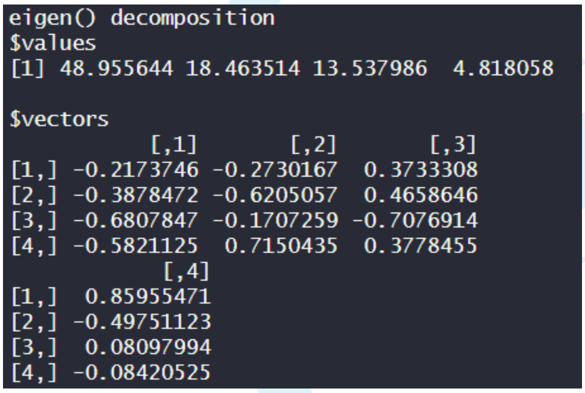
We begin with the principal component analysis for each dataset (men and women). To do this, we perform the decomposition of the covariance matrix S into its eigenvectors and eigenvalues:



```
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
```

Figure 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
```

Figure 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:

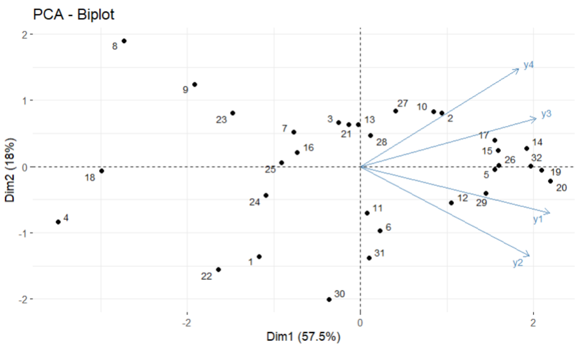


Figure 3: PCA for men

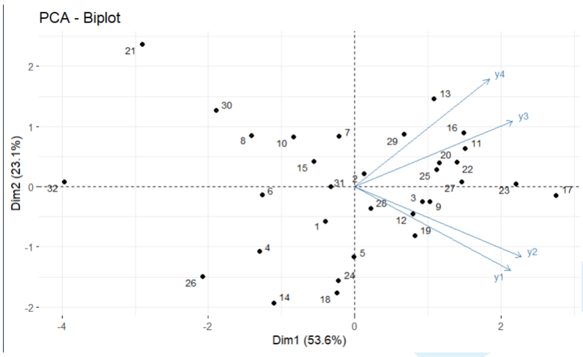


Figure 4: PCA for women

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.

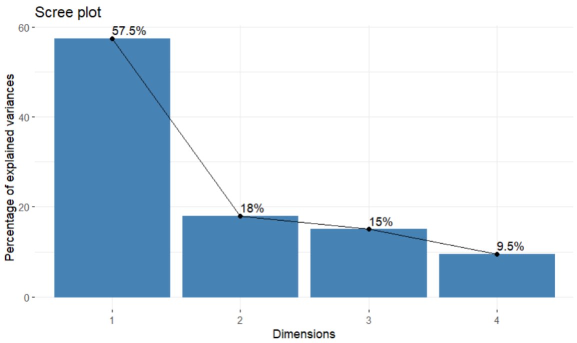


Figure 5: Elbow method for men

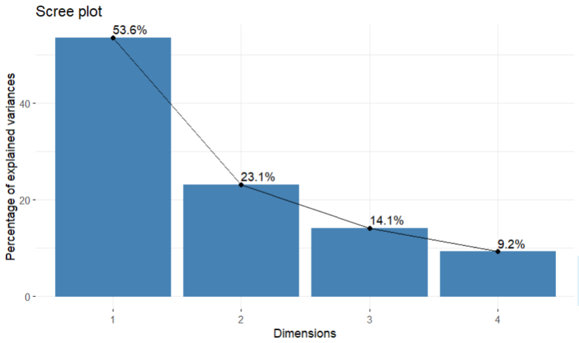


Figure 6: Elbow method for women

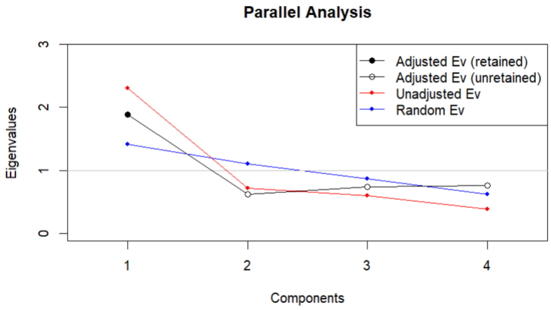


Figure 7: Parallel method for men

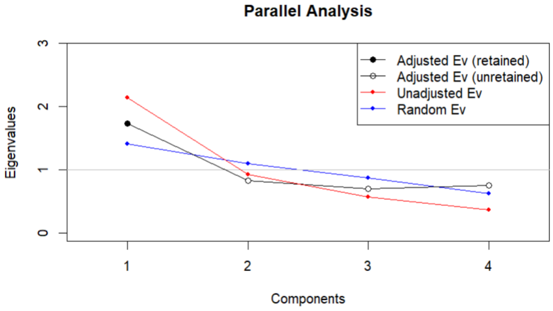


Figure 8: Parallel method for women

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:

- $y3$ and $y4$ are positively correlated with each other and have a positive correlation with Dim1.
- $y1$ and $y2$ 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

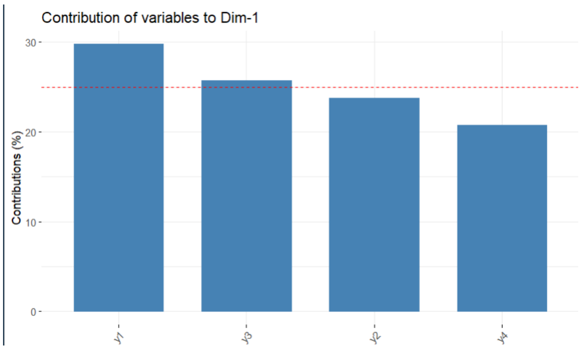


Figure 9: Contribution of each variable in the first principal component for men

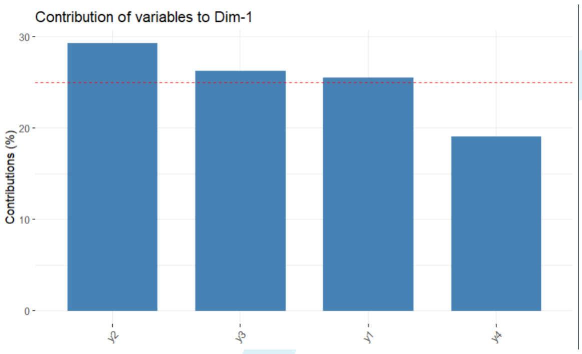
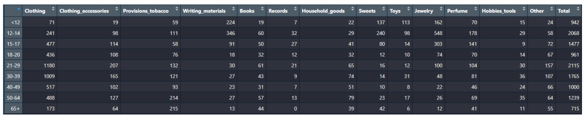


Figure 10: Contribution of each variable in the first principal component for women

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.

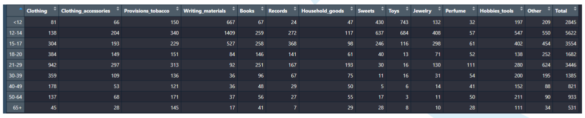
2 Correspondence Analysis

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



	Clothing	Clothing_accessories	Provisions_tobacco	Writing_materials	Books	Records	Household_goods	Sweets	Toys	Jewelry	Perfume	Hobbies_tools	Other	Total
<12	71	19	59	224	19	7	22	137	113	162	70	15	24	942
12-14	241	98	111	346	60	32	29	240	98	548	178	29	58	2068
15-17	477	114	58	91	50	27	41	80	14	303	141	9	72	1477
18-20	436	108	76	18	32	12	32	12	10	74	70	14	67	961
21-29	1180	207	132	30	61	21	65	16	12	100	104	30	157	2115
30-39	1009	165	121	27	43	9	74	14	31	48	81	36	107	1765
40-49	517	102	93	23	31	7	51	10	8	22	46	24	66	1000
50-64	488	127	214	27	57	13	79	23	17	26	69	35	64	1239
65+	173	64	215	13	44	0	39	42	6	12	41	11	55	715

Figure 11: Contingency table for womens



	Clothing	Clothing_accessories	Provisions_tobacco	Writing_materials	Books	Records	Household_goods	Sweets	Toys	Jewelry	Perfume	Hobbies_tools	Other	Total
<12	81	66	150	667	67	24	47	430	743	132	32	197	209	2845
12-14	138	204	340	1409	259	272	117	637	684	408	57	547	550	5622
15-17	304	193	229	527	258	368	98	246	116	298	61	402	454	3554
18-20	384	149	151	84	146	141	61	40	13	71	52	138	252	1682
21-29	942	297	313	92	251	167	193	30	16	130	111	280	624	3446
30-39	359	109	136	36	96	67	75	11	16	31	54	200	195	1385
40-49	178	53	121	36	48	29	50	5	6	14	41	152	88	821
50-64	137	68	171	37	56	27	55	17	3	11	50	211	90	933
65+	45	28	145	17	41	7	29	28	8	10	28	111	34	531

Figure 12: Contingency table for men

2.1 Correspondence analysis for women

How many dimensions to consider:

The extracted dimensions explain the variance in the data. Dimension 1 explains 76.18% of the variance, followed by Dimension 2 with 15.07%. Together, the first two dimensions explain 91.25% of the total variance in the data, and since it is recommended to select a sufficient number of dimensions that capture a significant amount of variance in the 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.

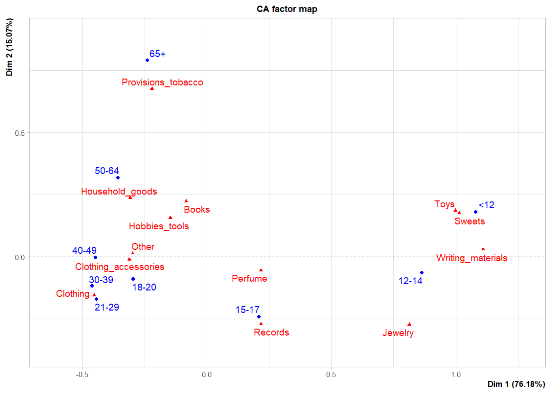


Figure 13: Correspondence analysis for women

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.

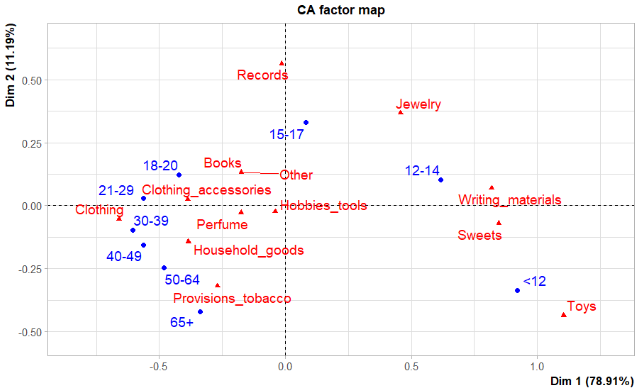


Figure 14: Correspondence analysis for men

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.

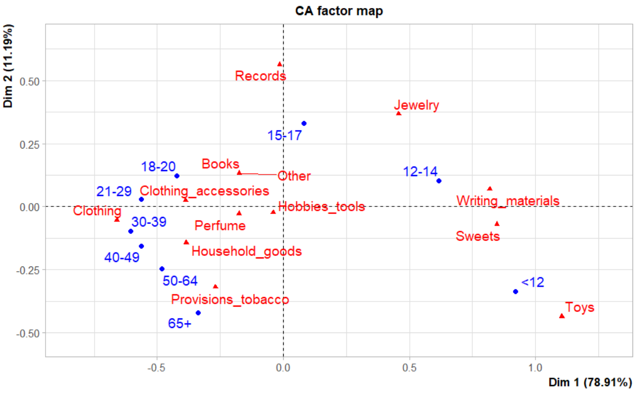


Figure 15: Joint correspondence analysis

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

References

Heijden, P. G. M., Falguerolles, A., & 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