Product Analysis Report

Prepared by

Allegra Adinolfi
Lucia La Forgia
Stefano Balla

1 Introduction

Since the spread of the online marketplace, people have become comfortable with the habit of buying clothing online. On this line, our survey nails at finding how this relates to the Luxury market in the Fashion Industry.

The Luxury market has always been oriented towards purchases in physical shops but, however to a lesser extent, lately, it has been influenced by a move towards online shopping. In particular, in the last three years and especially during 2020 Pandemic, the online marketplace saw an unchallenged increase.

According to Italia Online, *"E-commerce represents a huge opportunity for the luxury sector. Not only has it enabled many companies to survive during the lockdown, but it is also a key channel for recovery. Globally, online sales of luxury goods will increase until 2025, partly replacing physical shops"*.

More specifically, the main goal of this survey is to understand if our audience is interested in shopping luxury goods and what are their preferences over its purchase methods, in particular if online distribution channels can become the most popular new frontier.

1.1 Survey Description

Our survey was carried out on Microsoft Forms so that it would be possible to easily sum up responses in an Excel file.

We collected 124 responses. The average time to complete the questionnaire is 3 minutes and 47 seconds.

The questionnaire consists of 23 questions, of which some are identifiable as behavioural or opinion questions. We also used a screening question and a few of socio-demographic nature.

For many questions it was necessary to use a ranking scale, we opted for a Likert scale of 1-7.

SOCIO-DEMOGRAPHIC QUESTIONS:

These questions permit us to describe the characteristics of our audience and represent our categorical variables, over which we will do our analysis.

In particular, we asked to our audience to:

1. *Indicate gender*

2. *Indicate age*

3. *Indicate profession*

4. *Indicate where currently living*

SCREENING QUESTION:

This is a single question over which we screen our responses between those we are interested in and the rest.

5. *How much are you interested in the luxury market concerning the Fashion Industry?*

BEHAVIOURAL QUESTIONS:

Depending on the question, the Likert scale has been interpreted as a classic agreement-scale (1 ⇒ completely disagree; 7 ⇒ completely agree) or as a behavioural frequency percentage (1 ⇒ 0%; 7 ⇒ 100%):

6. *In order to buy luxury products, do you prefer physical shops [1] or online stores [7]?*

7. *In order to buy generic goods, do you usually rely on physical shops [1] or online shops [7]?*

8. *In order to buy accessories (bags, belts, shoes, etc.), select if you mostly purchase in physical stores [1] or online [7].*

9. *In order to buy clothes, select if you mostly purchase in physical stores [1] or online [7].*

10. *Select how much of your clothing shoppings **in physical stores** concerns luxury brands.*

11. *Select how much of your clothes **online** shoppings concerns a luxury brand.*

12. *By relying on shopping online, do you have the possibility to buy from brands not easily reachable otherwise?*

13. *In order to find discounts or promotions, select if you rely mostly on physical shops [1], like Outlets, or online shops [7].*

OPINION QUESTIONS:

These questions represent our **golden questions** and asked to our audience to specify how much each respondent agreed or disagreed with the statements, over a scale 1-7. The responses to these questions will all translate into quantitative variables.

Here is the list of our golden questions:

14. *Buying online is the same as buying in physical stores.*

15. *When I buy online, I only buy from the brand’s official website.*

16. *I prefer to buy online because it saves me time.*

17. *Buying in a physical store allows me to support the community I live in.*

18. *Buying online allows me to find specific products (limited editions, collaborations, etc.) that I would not find elsewhere.*

19. *I prefer to buy in physical shops to try on the clothes that I wish to buy.*

20. *The presence of discounts affects my willingness to buy.*

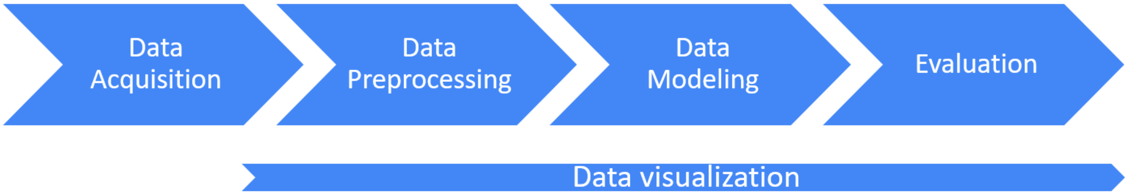
21. *I do not trust paying a large amount of money online.*

22. *In order to buy online, it is essential to have prior knowledge about the fit and size of the brand.*

23. *I prefer to buy in physical stores because I can get advice.*

1.2 Data Analysis

As previously said, our focus was to understand whether our audience is more oriented towards shopping online or shopping in physical shops when it comes to Luxury shopping. In order to do so, we followed the classical pipeline consisting in:



DATA ACQUISITION:

We collected our data from the responses of our survey.

The questionnaire was spread through social networks and personal contacts. We tried to reach an audience as heterogeneous as possible.

At the end of this step, we had gained 124 submissions. The data was stored in a single Excel file.

DATA PREPROCESSING:

This step was crucial to prepare our data in order to get it through the algorithms of the modelling phase. We got rid of the inconsistent lines of our dataset and we used our screening question to recognize useful and representative samples in our dataset. We also carried out size effect analysis and dimensionality reduction.

DATA MODELLING:

We performed clusterization over our dataset and executed chi-square analysis on its different categorical variables.

EVALUATION:

Thanks to the previous steps, we have been able to describe the characteristics of our clusters.

2 Preliminary analysis

2.1 SAS Library and Variables' names and labels

We performed our algorithms through the SAS System on a Windows 10 Virtual Machine.

The answers of the survey, as said before, were collected into a Microsoft Excel sheet and then imported in our library on the SAS program.

We named our dataset "*Luxury*". We performed a renaming to the columns of our dataset in order to handle them easily through the code, setting names from *d1* to *d29* (they are more than 23 because of extra variables automatically saved by the Form-to-Excel converter, such as "*Ora_di_inizio*" and "*Ora_di_completamento*" of the response process).

Concerning the columns' labels, we set short-phrase-plus-number labels in a way that made it easy for us to identify the relative question on the survey on the label basis, with no attempt to make them user-friendly, since it was not important for the final considerations on the clusters.

2.2 Data Cleaning

To prepare our dataset, we went through some steps:

- since we found just a few NULL values, we opted for removing the lines with this problem
- we also dropped the columns that were not relevant to our case (*Ora_di_inizio*, *Ora_di_completamento*, etc.)
- according to the scope of our survey, we designed a threshold to filter the dataset's rows on the basis of the screening question, aiming to consider only those interested in the Luxury market in the Fashion industry. We defined value 3 as our threshold.

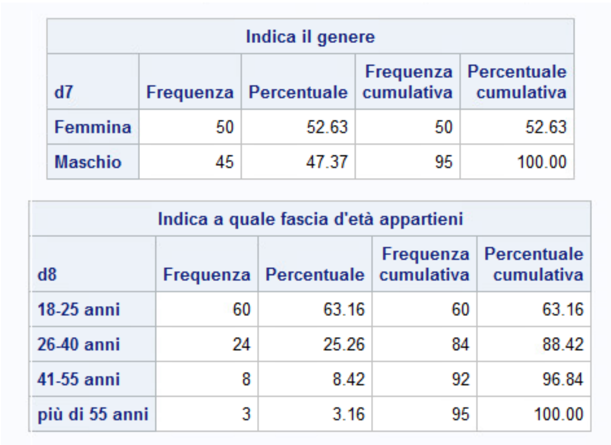
At the end of this process, *Luxury* dataset had 95 rows and the columns representing the informative part of the dataset, some of which were categorical variables and others were numerical ones.

2.3 Frequency Procedure

During the preliminary analysis, the first procedure to be used was the ***proc freq***, to show the frequency of the dataset's columns after the cleaning process in order to have a look at its basic information.

Thanks to this procedure, we were able to verify if the dataset was balanced or not considering our socio-demographic questions and check the distribution of the answers through each variable.
This kind of primary classification is crucial, because unbalanced data could affect the final results.

As shown in the following tables, the dataset is balanced according to the *gender*, while it is not considering *age*, *profession* and *geographic origin*. Surely, the unbalance of the dataset is due to our inability to reach different kinds of testers during the data collection phase and, as we will see during Chi-square analysis, this leak of information limited us from making some further considerations.

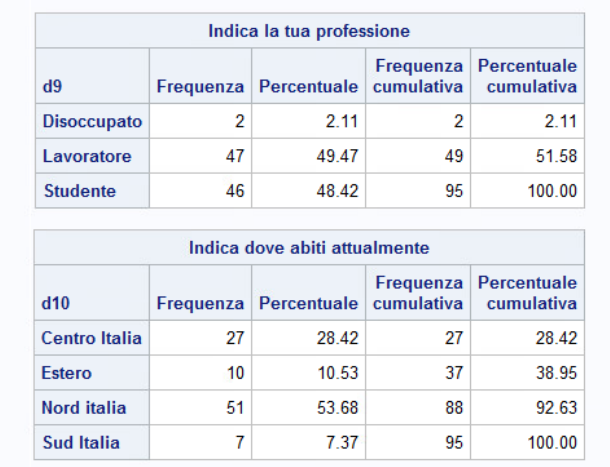


Indica il genere

d7	Frequenza	Percentuale	Frequenza cumulativa	Percentuale cumulativa
Femmina	50	52.63	50	52.63
Maschio	45	47.37	95	100.00

Indica a quale fascia d'età appartieni

d8	Frequenza	Percentuale	Frequenza cumulativa	Percentuale cumulativa
18-25 anni	60	63.16	60	63.16
26-40 anni	24	25.26	84	88.42
41-55 anni	8	8.42	92	96.84
più di 55 anni	3	3.16	95	100.00



Indica la tua professione

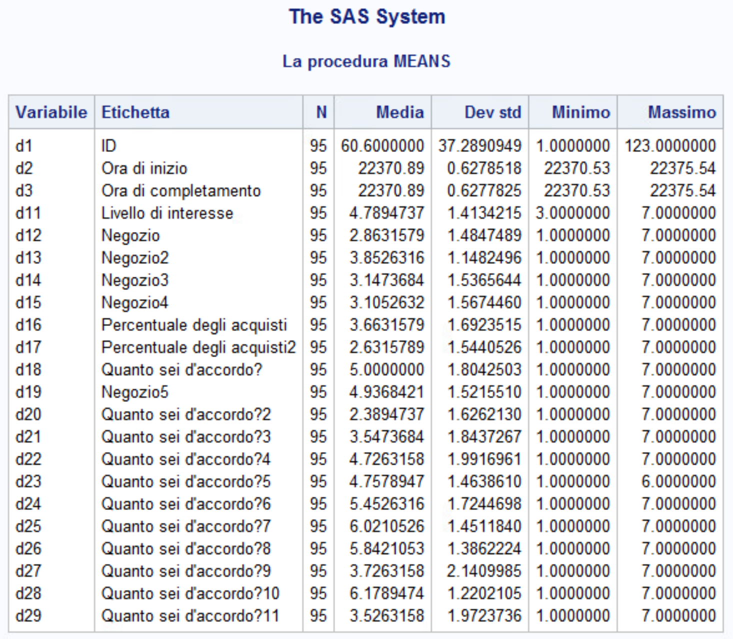
d9	Frequenza	Percentuale	Frequenza cumulativa	Percentuale cumulativa
Disoccupato	2	2.11	2	2.11
Lavoratore	47	49.47	49	51.58
Studente	46	48.42	95	100.00

Indica dove abiti attualmente

d10	Frequenza	Percentuale	Frequenza cumulativa	Percentuale cumulativa
Centro Italia	27	28.42	27	28.42
Estero	10	10.53	37	38.95
Nord italia	51	53.68	88	92.63
Sud Italia	7	7.37	95	100.00

2.4 Means Procedure

In addition to the previous procedure, we used ***proc means*** to show intrinsic properties of the dataset:



The SAS System

La procedura MEANS

Variabile	Etichetta	N	Media	Dev std	Minimo	Massimo
d1	ID	95	60.6000000	37.2890949	1.0000000	123.0000000
d2	Ora di inizio	95	22370.89	0.6278518	22370.53	22375.54
d3	Ora di completamento	95	22370.89	0.6277825	22370.53	22375.54
d11	Livello di interesse	95	4.7894737	1.4134215	3.0000000	7.0000000
d12	Negozio	95	2.8631579	1.4847489	1.0000000	7.0000000
d13	Negozio2	95	3.8526316	1.1482496	1.0000000	7.0000000
d14	Negozio3	95	3.1473684	1.5365644	1.0000000	7.0000000
d15	Negozio4	95	3.1052632	1.5674460	1.0000000	7.0000000
d16	Percentuale degli acquisti	95	3.6631579	1.6923515	1.0000000	7.0000000
d17	Percentuale degli acquisti2	95	2.6315789	1.5440526	1.0000000	7.0000000
d18	Quanto sei d'accordo?	95	5.0000000	1.8042503	1.0000000	7.0000000
d19	Negozio5	95	4.9368421	1.5215510	1.0000000	7.0000000
d20	Quanto sei d'accordo?2	95	2.3894737	1.6262130	1.0000000	7.0000000
d21	Quanto sei d'accordo?3	95	3.5473684	1.8437267	1.0000000	7.0000000
d22	Quanto sei d'accordo?4	95	4.7263158	1.9916961	1.0000000	7.0000000
d23	Quanto sei d'accordo?5	95	4.7578947	1.4638610	1.0000000	6.0000000
d24	Quanto sei d'accordo?6	95	5.4526316	1.7244698	1.0000000	7.0000000
d25	Quanto sei d'accordo?7	95	6.0210526	1.4511840	1.0000000	7.0000000
d26	Quanto sei d'accordo?8	95	5.8421053	1.3862224	1.0000000	7.0000000
d27	Quanto sei d'accordo?9	95	3.7263158	2.1409985	1.0000000	7.0000000
d28	Quanto sei d'accordo?10	95	6.1789474	1.2202105	1.0000000	7.0000000
d29	Quanto sei d'accordo?11	95	3.5263158	1.9723736	1.0000000	7.0000000

As we can see from the image above, this procedure allows the user to find the average, the minimum and the maximum values for each variable of the dataset, plus its standard deviation.

The Means Procedure also gives the opportunity for further insights, as those following. In order to understand them, we recall that we used a Likert scale 1-7.

We notice that the maximum value of question *d23* is 6, meaning that none of the respondents selected the maximum value of 7. At first sight, this might appear to be due to a skewed distribution of the variable towards its minimum value, but its average tends to the median value of the distribution, with a small enough standard deviation, meaning that, conversely to the first impression, the variable is distributed mainly around the central value.

Furthermore, we point out that, for question *d11*, the minimum value is 3, this is because we dropped all the rows where this value was lower than this threshold, in order to continue with our analysis only with those lines that declared themselves sufficiently interested in the luxury market.

3 Principal component analysis

In this section, we explore our dataset in order to understand if it is subject to the size effect and if there is the need to pursue to reduce it. Then, we will discuss dimensionality reduction, by means of PCA.

3.1 Size effect: recognition and elimination

The size effect is a common phenomenon that may occur when analyzing behavioral or opinion data. This is generated by the tendency of each respondent to use the given Likert scale in an internal and personal way, inconsistently with its natural, suitable distribution.

To identify the presence of size effect we checked for three aspects:

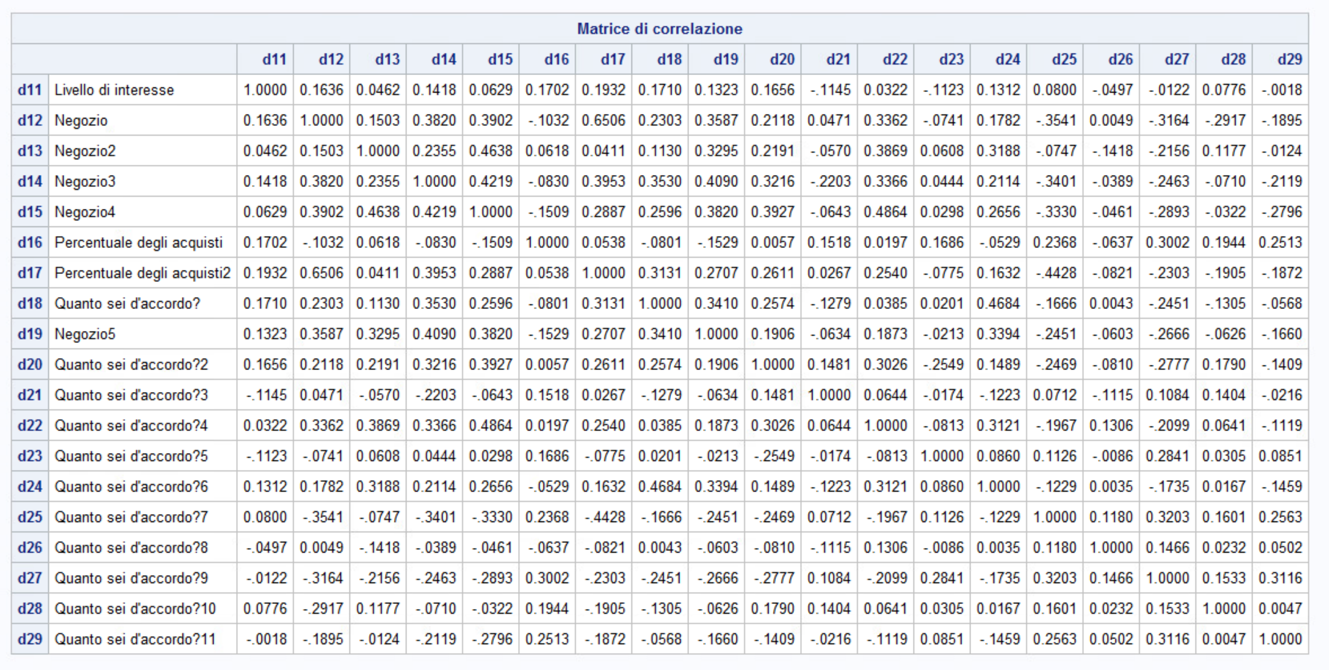
1. a general positive trend in the correlation matrix;
2. positive values in the first Principal Component;
3. a correlation coefficient between the average of each variable and PC1 that tend to 1.

General positive trend in the correlation matrix:

To give an answer to this point, we plotted the correlation matrix of the variables, reported in the following table.

Since the matrix is symmetric, we can consider only the values in the bottom part under the diagonal. What we observed is that there is no strong positive trend, indeed we find a lot of negative values. However, we have decided to proceed with the analysis of size effect.

In addition, it is important to highlight the absence of very high values, in fact the maximum correlation value is the one between the couple d12-d17, with a value of approximately 65%, and, in general, there is a low presence of values higher than 50%. That means the questions are independent to each other, therefore all of them carry different, non-reducible information. For this reason, we can make the assumption that the questions were well-written.



Matrice di correlazione

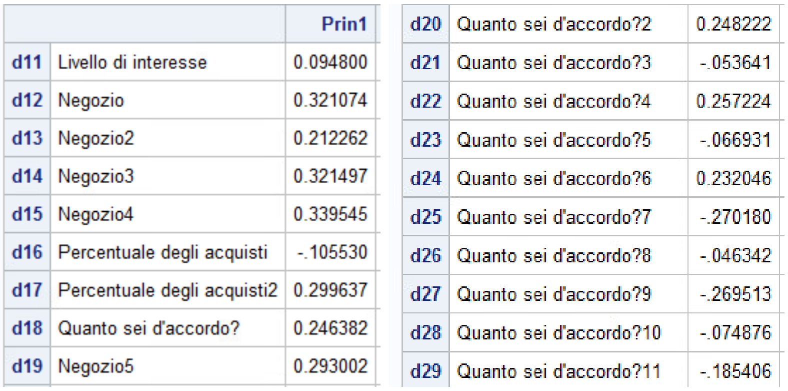
		d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22	d23	d24	d25	d26	d27	d28	d29
d11	Livello di interesse	1.0000	0.1636	0.0462	0.1418	0.0629	0.1702	0.1932	0.1710	0.1323	0.1656	-.1145	0.0322	-.1123	0.1312	0.0800	-.0497	-.0122	0.0776	-.0018
d12	Negozio	0.1636	1.0000	0.1503	0.3820	0.3902	-.1032	0.6506	0.2303	0.3587	0.2118	0.0471	0.3362	-.0741	0.1782	-.3541	0.0049	-.3164	-.2917	-.1895
d13	Negozio2	0.0462	0.1503	1.0000	0.2355	0.4638	0.0618	0.0411	0.1130	0.3295	0.2191	-.0570	0.3869	0.0608	0.3188	-.0747	-.1418	-.2156	0.1177	-.0124
d14	Negozio3	0.1418	0.3820	0.2355	1.0000	0.4219	-.0830	0.3953	0.3530	0.4090	0.3216	-.2203	0.3366	0.0444	0.2114	-.3401	-.0389	-.2463	-.0710	-.2119
d15	Negozio4	0.0629	0.3902	0.4638	0.4219	1.0000	-.1509	0.2887	0.2596	0.3820	0.3927	-.0643	0.4864	0.0298	0.2656	-.3330	-.0461	-.2893	-.0322	-.2796
d16	Percentuale degli acquisti	0.1702	-.1032	0.0618	-.0830	-.1509	1.0000	0.0538	-.0801	-.1529	0.0057	0.1518	0.0197	0.1686	-.0529	0.2368	-.0637	0.3002	0.1944	0.2513
d17	Percentuale degli acquisti2	0.1932	0.6506	0.0411	0.3953	0.2887	0.0538	1.0000	0.3131	0.2707	0.2611	0.0267	0.2540	-.0775	0.1632	-.4428	-.0821	-.2303	-.1905	-.1872
d18	Quanto sei d'accordo?	0.1710	0.2303	0.1130	0.3530	0.2596	-.0801	0.3131	1.0000	0.3410	0.2574	-.1279	0.0385	0.0201	0.4684	-.1666	0.0043	-.2451	-.1305	-.0568
d19	Negozio5	0.1323	0.3587	0.3295	0.4090	0.3820	-.1529	0.2707	0.3410	1.0000	0.1906	-.0634	0.1873	-.0213	0.3394	-.2451	-.0603	-.2666	-.0626	-.1660
d20	Quanto sei d'accordo?2	0.1656	0.2118	0.2191	0.3216	0.3927	0.0057	0.2611	0.2574	0.1906	1.0000	0.1481	0.3026	-.2549	0.1489	-.2469	-.0810	-.2777	0.1790	-.1409
d21	Quanto sei d'accordo?3	-.1145	0.0471	-.0570	-.2203	-.0643	0.1518	0.0267	-.1279	-.0634	0.1481	1.0000	0.0644	-.0174	-.1223	0.0712	-.1115	0.1084	0.1404	-.0216
d22	Quanto sei d'accordo?4	0.0322	0.3362	0.3869	0.3366	0.4864	0.0197	0.2540	0.0385	0.1873	0.3026	0.0644	1.0000	-.0813	0.3121	-.1967	0.1306	-.2099	0.0641	-.1119
d23	Quanto sei d'accordo?5	-.1123	-.0741	0.0608	0.0444	0.0298	0.1686	-.0775	0.0201	-.0213	-.2549	-.0174	-.0813	1.0000	0.0860	0.1126	-.0086	0.2841	0.0305	0.0851
d24	Quanto sei d'accordo?6	0.1312	0.1782	0.3188	0.2114	0.2656	-.0529	0.1632	0.4684	0.3394	0.1489	-.1223	0.3121	0.0860	1.0000	-.1229	0.0035	-.1735	0.0167	-.1459
d25	Quanto sei d'accordo?7	0.0800	-.3541	-.0747	-.3401	-.3330	0.2368	-.4428	-.1666	-.2451	-.2469	0.0712	-.1967	0.1126	-.1229	1.0000	0.1180	0.3203	0.1601	0.2563
d26	Quanto sei d'accordo?8	-.0497	0.0049	-.1418	-.0389	-.0461	-.0637	-.0821	0.0043	-.0603	-.0810	-.1115	0.1306	-.0086	0.0035	0.1180	1.0000	0.1466	0.0232	0.0502
d27	Quanto sei d'accordo?9	-.0122	-.3164	-.2156	-.2463	-.2893	0.3002	-.2303	-.2451	-.2666	-.2777	0.1084	-.2099	0.2841	-.1735	0.3203	0.1466	1.0000	0.1533	0.3116
d28	Quanto sei d'accordo?10	0.0776	-.2917	0.1177	-.0710	-.0322	0.1944	-.1905	-.1305	-.0626	0.1790	0.1404	0.0641	0.0305	0.0167	0.1601	0.0232	0.1533	1.0000	0.0047
d29	Quanto sei d'accordo?11	-.0018	-.1895	-.0124	-.2119	-.2796	0.2513	-.1872	-.0568	-.1660	-.1409	-.0216	-.1119	0.0851	-.1459	0.2563	0.0502	0.3116	0.0047	1.0000

Presence of positive values in the first Principal Component:

Using PCA, we are able to represent the whole dataset in new dimensions called Principal Components. Moreover, we can decide to keep only some of these dimensions in order to reduce the dimensionality of the dataset.

In this case, we select only the First Principal Component, that means we are summarizing all the dataset with a single dimension. In particular, we look for a skewed behaviour along this axis, because it would be a signal for the size effect.

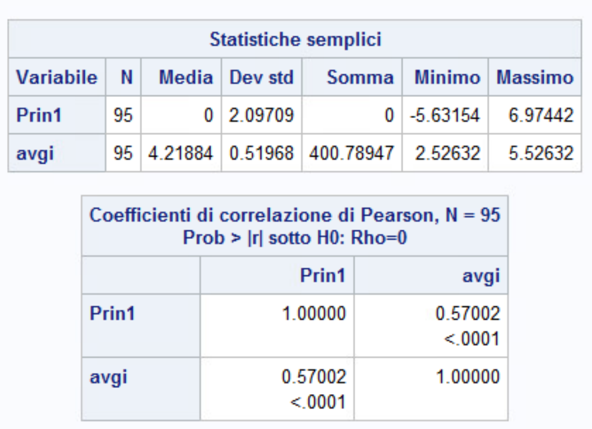
As we can see from the following image, the variables are well distributed over the first Principal Component, so we can say we have no clear signal of the size effect as we would have stated if all the values were positive. This is another way to see that the correlation matrix is not completely positive and also to see that PC1 mainly describes the variables d12, d14 and d15.



		Prin1
d11	Livello di interesse	0.094800
d12	Negozio	0.321074
d13	Negozio2	0.212262
d14	Negozio3	0.321497
d15	Negozio4	0.339545
d16	Percentuale degli acquisti	-.105530
d17	Percentuale degli acquisti2	0.299637
d18	Quanto sei d'accordo?	0.246382
d19	Negozio5	0.293002
d20	Quanto sei d'accordo?2	0.248222
d21	Quanto sei d'accordo?3	-.053641
d22	Quanto sei d'accordo?4	0.257224
d23	Quanto sei d'accordo?5	-.066931
d24	Quanto sei d'accordo?6	0.232046
d25	Quanto sei d'accordo?7	-.270180
d26	Quanto sei d'accordo?8	-.046342
d27	Quanto sei d'accordo?9	-.269513
d28	Quanto sei d'accordo?10	-.074876
d29	Quanto sei d'accordo?11	-.185406

Correlation coefficient between variables' average and PC1 close to 1:

The last signal to the size effect is suggested by the correlation between the average of the answers of all the respondents for each variable and the PC1. A value that tends towards 1 indicates that all the information contained in the dataset is well represented by the PC1.



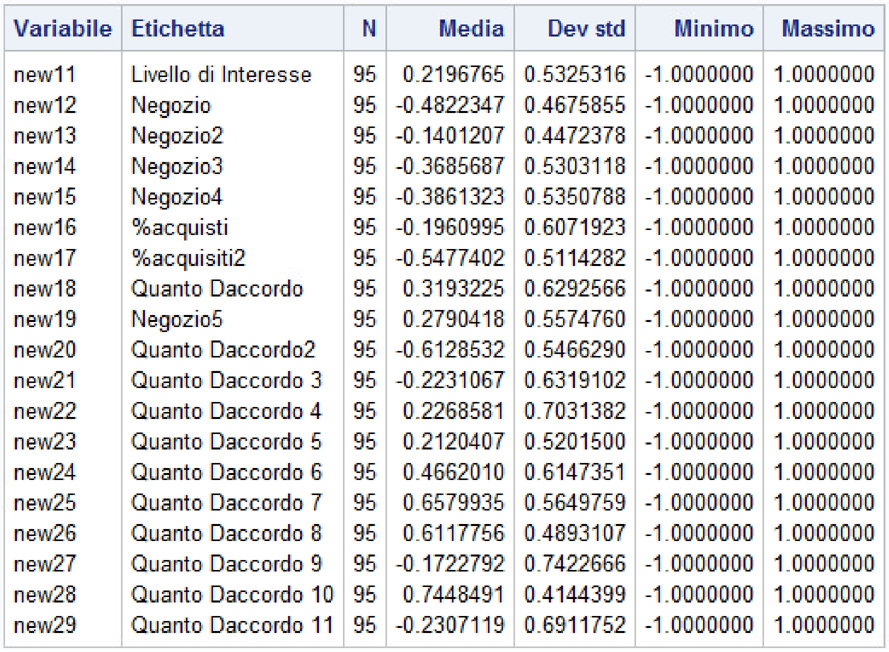
Statistiche semplici

Variabile	N	Media	Dev std	Somma	Minimo	Massimo
Prin1	95	0	2.09709	0	-5.63154	6.97442
avgi	95	4.21884	0.51968	400.78947	2.52632	5.52632

Coefficienti di correlazione di Pearson, N = 95
Prob > \|r\| sotto H0: Rho=0

	Prin1	avgi
Prin1	1.00000	0.57002 <.0001
avgi	0.57002 <.0001	1.00000

The correlation value between PC1 and the average is high enough to indicate the possibility of having the size effect (r=0.57) and, since the Principal Component Analysis is affected by scaling, we also rescale the dataset by the personal average. Doing it, we have obtained a new rescaled dataset that is now free from the size effect. In the following table, the results of *proc means* over the rescaled dataset:



Variabile	Etichetta	N	Media	Dev std	Minimo	Massimo
new11	Livello di Interesse	95	0.2196765	0.5325316	-1.0000000	1.0000000
new12	Negozio	95	-0.4822347	0.4675855	-1.0000000	1.0000000
new13	Negozio2	95	-0.1401207	0.4472378	-1.0000000	1.0000000
new14	Negozio3	95	-0.3685687	0.5303118	-1.0000000	1.0000000
new15	Negozio4	95	-0.3861323	0.5350788	-1.0000000	1.0000000
new16	%acquisti	95	-0.1960995	0.6071923	-1.0000000	1.0000000
new17	%acquisiti2	95	-0.5477402	0.5114282	-1.0000000	1.0000000
new18	Quanto Daccordo	95	0.3193225	0.6292566	-1.0000000	1.0000000
new19	Negozio5	95	0.2790418	0.5574760	-1.0000000	1.0000000
new20	Quanto Daccordo2	95	-0.6128532	0.5466290	-1.0000000	1.0000000
new21	Quanto Daccordo 3	95	-0.2231067	0.6319102	-1.0000000	1.0000000
new22	Quanto Daccordo 4	95	0.2268581	0.7031382	-1.0000000	1.0000000
new23	Quanto Daccordo 5	95	0.2120407	0.5201500	-1.0000000	1.0000000
new24	Quanto Daccordo 6	95	0.4662010	0.6147351	-1.0000000	1.0000000
new25	Quanto Daccordo 7	95	0.6579935	0.5649759	-1.0000000	1.0000000
new26	Quanto Daccordo 8	95	0.6117756	0.4893107	-1.0000000	1.0000000
new27	Quanto Daccordo 9	95	-0.1722792	0.7422666	-1.0000000	1.0000000
new28	Quanto Daccordo 10	95	0.7448491	0.4144399	-1.0000000	1.0000000
new29	Quanto Daccordo 11	95	-0.2307119	0.6911752	-1.0000000	1.0000000

The new dataset maintains the original information, plus it is not affected by the scale problem anymore, so we will use it for the following techniques.

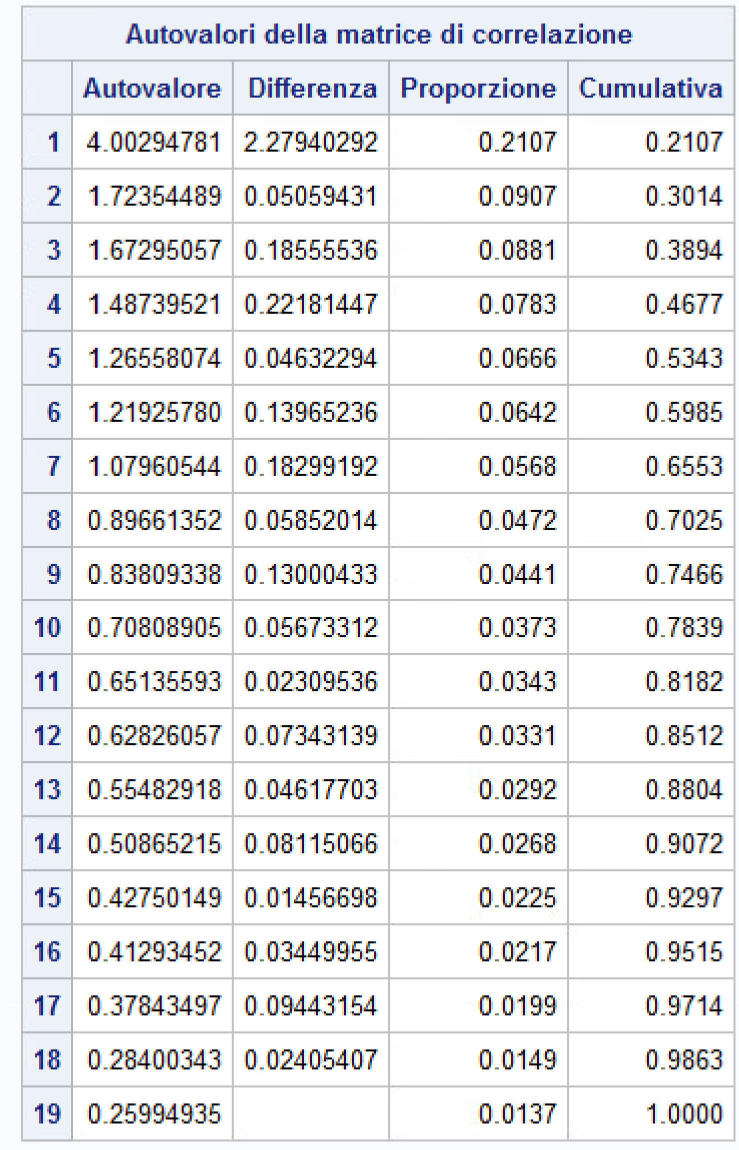
3.2 Dimensionality reduction

Dimensionality reduction has to be discussed by means of PCA and it will allow us to consider a reduced number of variables when creating clusters.

Multidimensionality is a hot topic in the field of data analysis because variables are often correlated with each other and therefore the information they bring is affected by noise and redundancy.

Through the use of PCA, it is possible to maintain the information in the dataset by discarding exclusively the background noise. This is performed by working over the correlation matrix and by decomposing it into eigenvalues and eigenvectors in order to translate the data on the new axes.
It is possible to reduce dimensionality by keeping a smaller number of dimensions so that to retain only a certain wealth of information, according to the magnitude of the eigenvalues. This amount of information is the variance ratio of the model, i.e. the informativeness that the new dataset retains from the original dataset.

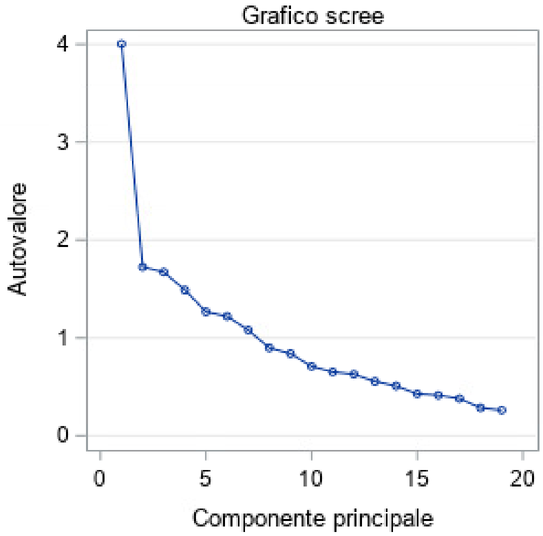
As already said, the starting point for these considerations are the eigenvalues:

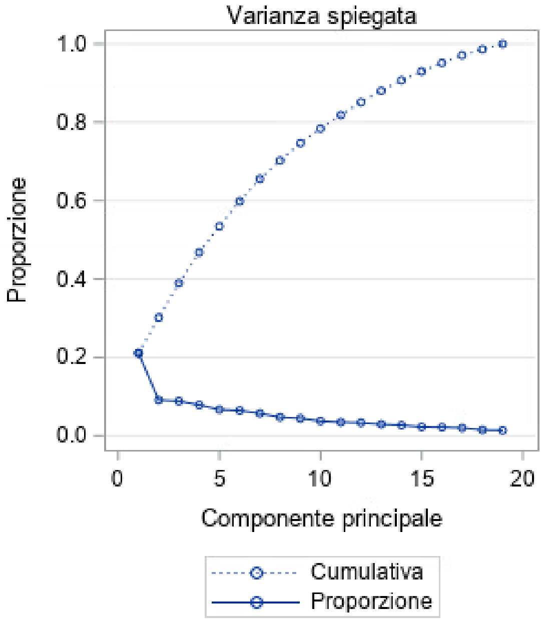


Autovalori della matrice di correlazione				
	Autovalore	Differenza	Proporzione	Cumulativa
1	4.00294781	2.27940292	0.2107	0.2107
2	1.72354489	0.05059431	0.0907	0.3014
3	1.67295057	0.18555536	0.0881	0.3894
4	1.48739521	0.22181447	0.0783	0.4677
5	1.26558074	0.04632294	0.0666	0.5343
6	1.21925780	0.13965236	0.0642	0.5985
7	1.07960544	0.18299192	0.0568	0.6553
8	0.89661352	0.05852014	0.0472	0.7025
9	0.83809338	0.13000433	0.0441	0.7466
10	0.70808905	0.05673312	0.0373	0.7839
11	0.65135593	0.02309536	0.0343	0.8182
12	0.62826057	0.07343139	0.0331	0.8512
13	0.55482918	0.04617703	0.0292	0.8804
14	0.50865215	0.08115066	0.0268	0.9072
15	0.42750149	0.01456698	0.0225	0.9297
16	0.41293452	0.03449955	0.0217	0.9515
17	0.37843497	0.09443154	0.0199	0.9714
18	0.28400343	0.02405407	0.0149	0.9863
19	0.25994935		0.0137	1.0000

At this point, it is crucial to choose the right number of dimensions to consider. There are several ways to do this:

- setting a threshold and searching along the cumulative percentage column for the number of dimensions that meet this threshold;
- looking at the percentage difference that each new dimension makes and, when a difference is very large, cutting the dimensions that come afterwards;
- use the elbow method by looking at the following plot:





Considering all the three methods described, we decided to use the first ten principal components, discarding those from 11 to 19, in order to maintain an explained variance percentage of 78%, which seemed a reasonable number to carry out the analysis.

In fact, to support this decision, we can see in the left graph above that, at the tenth dimension, there is a change in slope, from which the trend becomes linear.

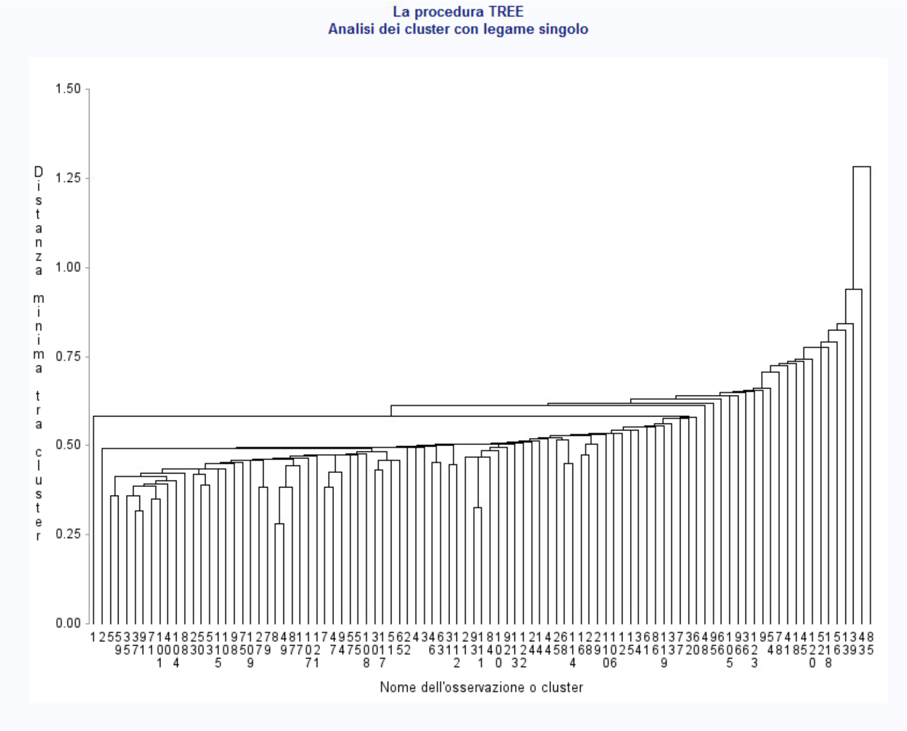
4 Clustering

This is the central phase of the analysis. In this chapter, we describe the clustering methods that we adopted, the criteria by which we identified the number of clusters and, finally, the salient characteristics of each cluster using some well-known tests, such as the **T-Test** for numerical variables and the **Chi-square** for categorical variables.

4.1 Clustering method

Since this is an unsupervised analysis, we adopted hierarchical clustering. This technique does not require the number of clusters to be specified a priori, as is the case of other types of unsupervised clustering, such as k-means clustering. Instead, this procedure generates clusters in a hierarchical manner, on the basis of the similarities of the samples.

In SAS, ***proc cluster*** uses, as default, the Euclidean distance to compare elements and investigate similarities. One parameter that must be included is the *method* parameter, which specifies the way clusters are generated starting from the distance matrix. We have tried two different methods: Single Linkage and Ward.
Here is the Single Linkage Cluster Analysis:

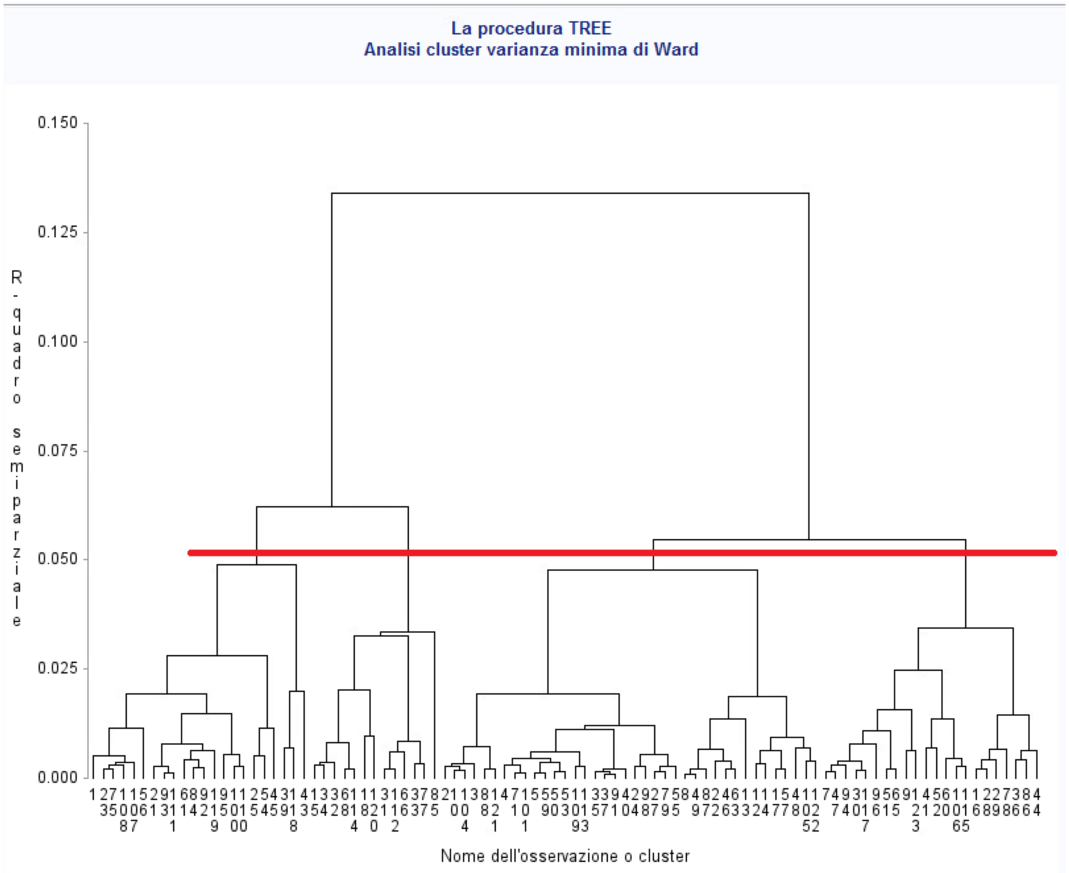


4.2 Cluster formation

By looking at the two Cluster Analysis trees, we decided to adopt the clusterization performed through Ward method because, in our case, it delivered the best result. This will be clear by comparing the previous tree, obtained by Single Linkage, with the one presented in the following section, by Ward method.

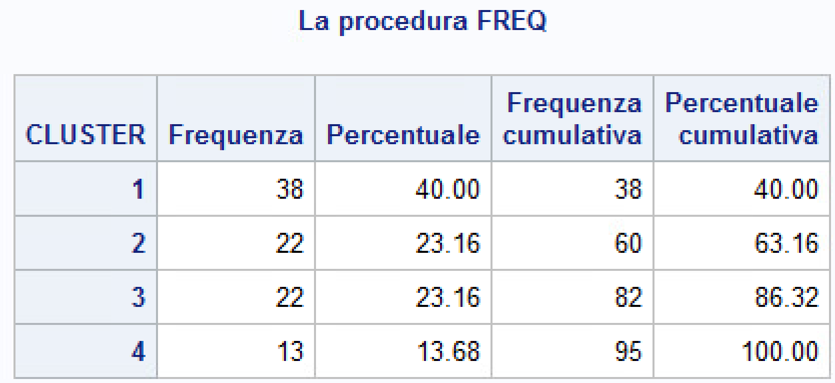
4.3 Dendrogram and cluster identification

The result of the *proc tree* is our Dendrogram, a tree over which different levels of aggregation are represented on the basis of the vertical axis, i.e. the similarity measure. The further up the tree one goes, the larger and more generic the clusters become, losing in specificity.



The number of clusters to be chosen depends on the pursued analysis. In our case, we decided to define four clusters.

The choice of this number is certainly influenced by the number of items available: by increasing the number of clusters, we would obtain groups with a very small number of samples, possibly encountering a less meaningful characterisation.

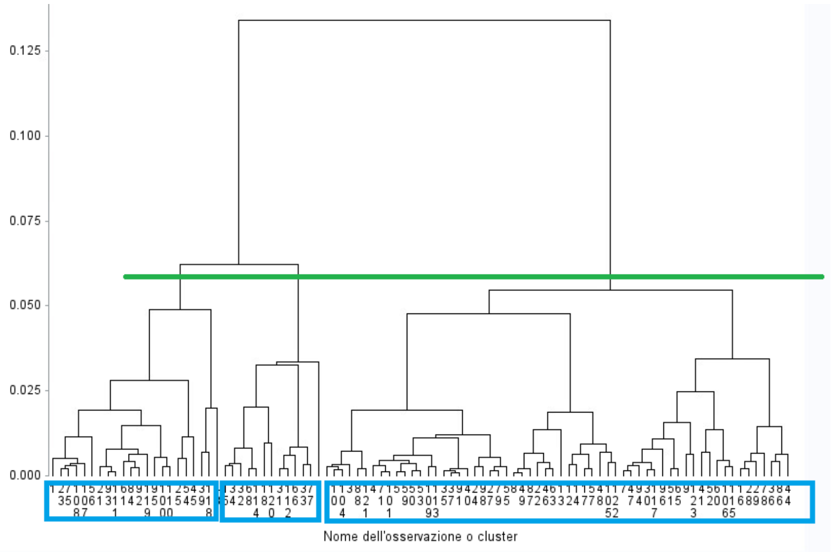


La procedura FREQ

CLUSTER	Frequenza	Percentuale	Frequenza cumulativa	Percentuale cumulativa
1	38	40.00	38	40.00
2	22	23.16	60	63.16
3	22	23.16	82	86.32
4	13	13.68	95	100.00

This image shows the result of ***proc freq*** on the cluster table.
From this, it can also be seen that the partitioning is not well balanced, as Cluster 1 has 40% of the respondents, making it the largest cluster by far. Cluster 2 and Cluster 3 have the same size, with 23% of the audience each, while Cluster 4, with 14% of the respondents, is the least populated.

However, reducing the number of clusters to three would not be beneficial, since it would not redistribute the audience more evenly among the clusters, but it would create a much larger cluster instead, as can be seen in the following tree.



4.4 T-test

In this section, the **t-test** is carried out to check which variables characterise each cluster. Each variable can define the cluster in a negative or positive way, this tendency is shown by the t-value. However, not all variables characterise the cluster, for this reason it is important to evaluate the **p-value**, that is the conditional probability of obtaining that value for that variable supposing the null hypothesis to be true. Therefore, it is necessary to set a threshold α, typically **α=0.05**, and compare the p-value with this threshold:

- if $p > \alpha$, the null hypothesis cannot be rejected, so the variable does not characterise the cluster;
- if $p < \alpha$, the null hypothesis is rejected, thus the variable characterises the cluster.

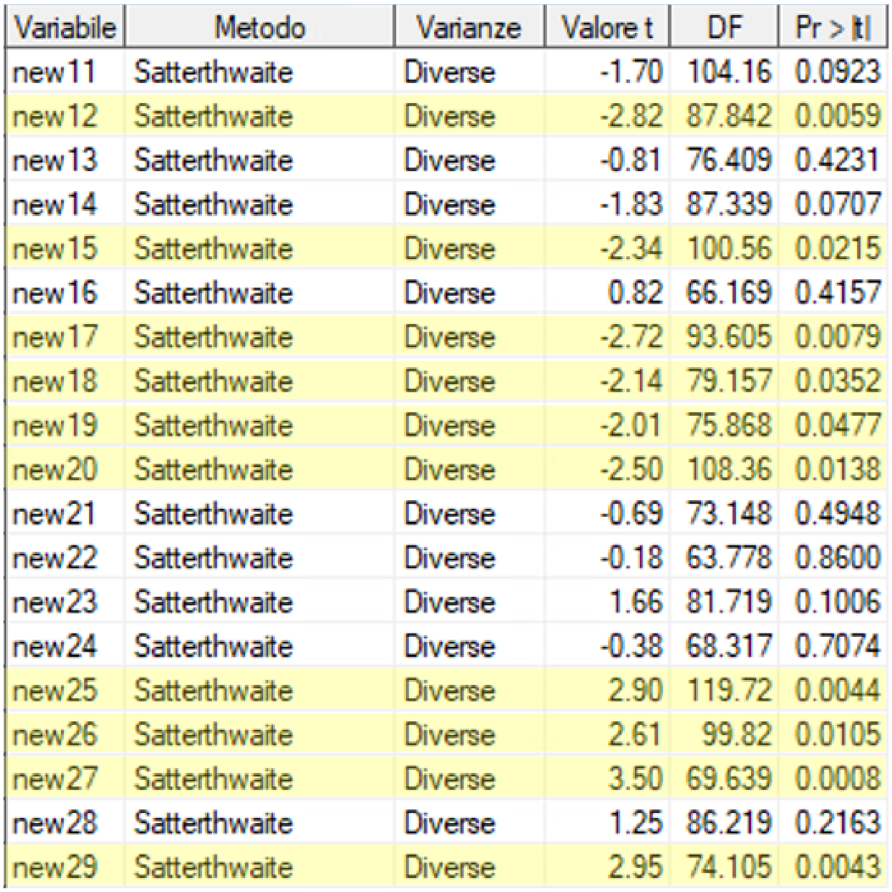
4.5 Cluster Analysis

The tables with the results of t-test are shown below for each cluster, highlighting the variables that meet the requirements.

CLUSTER 1

As can be seen from the table, the variables that meet the threshold-limit are ten: d12, d15, d17, d18, d19,d20, d25, d26, d27, d29.

At this stage we are not concerned with naming the variables, as we are only identifying what variables are meaningful for the definition of each cluster and in what way: positively or negatively.
The semantic meaning will be described later on and the important variables will be given an understandable name.

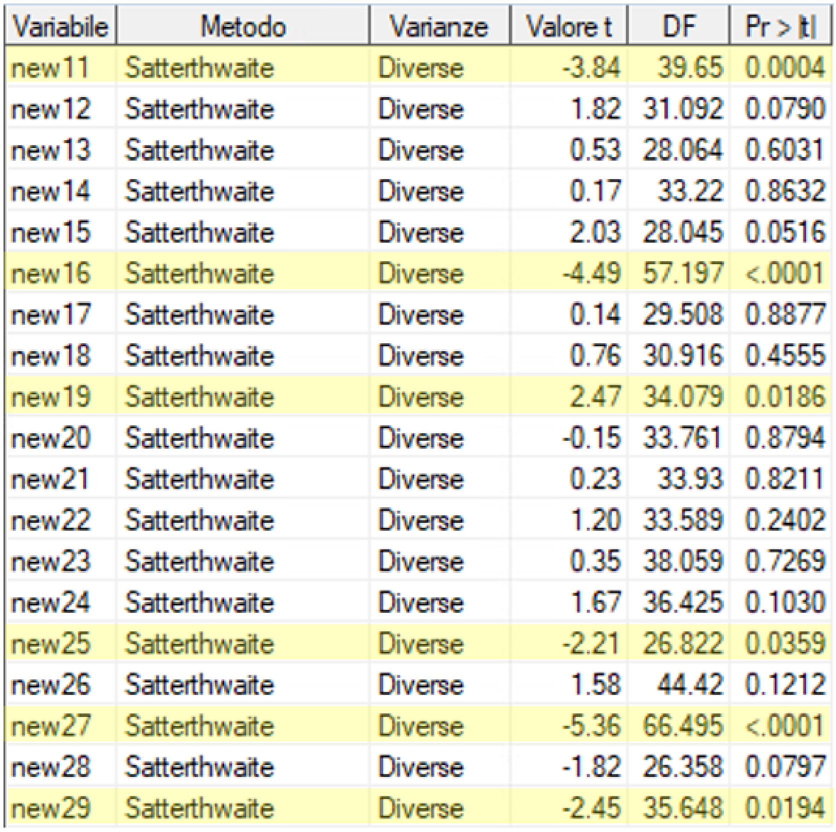


Variabile	Metodo	Varianze	Valore t	DF	Pr > |t|
new11	Satterthwaite	Diverse	-1.70	104.16	0.0923
new12	Satterthwaite	Diverse	-2.82	87.842	0.0059
new13	Satterthwaite	Diverse	-0.81	76.409	0.4231
new14	Satterthwaite	Diverse	-1.83	87.339	0.0707
new15	Satterthwaite	Diverse	-2.34	100.56	0.0215
new16	Satterthwaite	Diverse	0.82	66.169	0.4157
new17	Satterthwaite	Diverse	-2.72	93.605	0.0079
new18	Satterthwaite	Diverse	-2.14	79.157	0.0352
new19	Satterthwaite	Diverse	-2.01	75.868	0.0477
new20	Satterthwaite	Diverse	-2.50	108.36	0.0138
new21	Satterthwaite	Diverse	-0.69	73.148	0.4948
new22	Satterthwaite	Diverse	-0.18	63.778	0.8600
new23	Satterthwaite	Diverse	1.66	81.719	0.1006
new24	Satterthwaite	Diverse	-0.38	68.317	0.7074
new25	Satterthwaite	Diverse	2.90	119.72	0.0044
new26	Satterthwaite	Diverse	2.61	99.82	0.0105
new27	Satterthwaite	Diverse	3.50	69.639	0.0008
new28	Satterthwaite	Diverse	1.25	86.219	0.2163
new29	Satterthwaite	Diverse	2.95	74.105	0.0043

Variables d12, d15, d17, d18, d19, d20 have similar negative values not close to 0 and, in particular, variable d12 is the one with the lowest value. On the other hand, variables d25, d26, d27, d29 have positive values, once again not close to 0. We note variable d27, the one of the highest value.

CLUSTER 2

As we have already mentioned, Cluster 2 is smaller in size than Cluster 1. In this case, there are seven variables that meet the imposed threshold of α: d11, d16, d19, d25, d27, d29.



Variabile	Metodo	Varianze	Valore t	DF	Pr > |t|
new11	Satterthwaite	Diverse	-3.84	39.65	0.0004
new12	Satterthwaite	Diverse	1.82	31.092	0.0790
new13	Satterthwaite	Diverse	0.53	28.064	0.6031
new14	Satterthwaite	Diverse	0.17	33.22	0.8632
new15	Satterthwaite	Diverse	2.03	28.045	0.0516
new16	Satterthwaite	Diverse	-4.49	57.197	<.0001
new17	Satterthwaite	Diverse	0.14	29.508	0.8877
new18	Satterthwaite	Diverse	0.76	30.916	0.4555
new19	Satterthwaite	Diverse	2.47	34.079	0.0186
new20	Satterthwaite	Diverse	-0.15	33.761	0.8794
new21	Satterthwaite	Diverse	0.23	33.93	0.8211
new22	Satterthwaite	Diverse	1.20	33.589	0.2402
new23	Satterthwaite	Diverse	0.35	38.059	0.7269
new24	Satterthwaite	Diverse	1.67	36.425	0.1030
new25	Satterthwaite	Diverse	-2.21	26.822	0.0359
new26	Satterthwaite	Diverse	1.58	44.42	0.1212
new27	Satterthwaite	Diverse	-5.36	66.495	<.0001
new28	Satterthwaite	Diverse	-1.82	26.358	0.0797
new29	Satterthwaite	Diverse	-2.45	35.648	0.0194

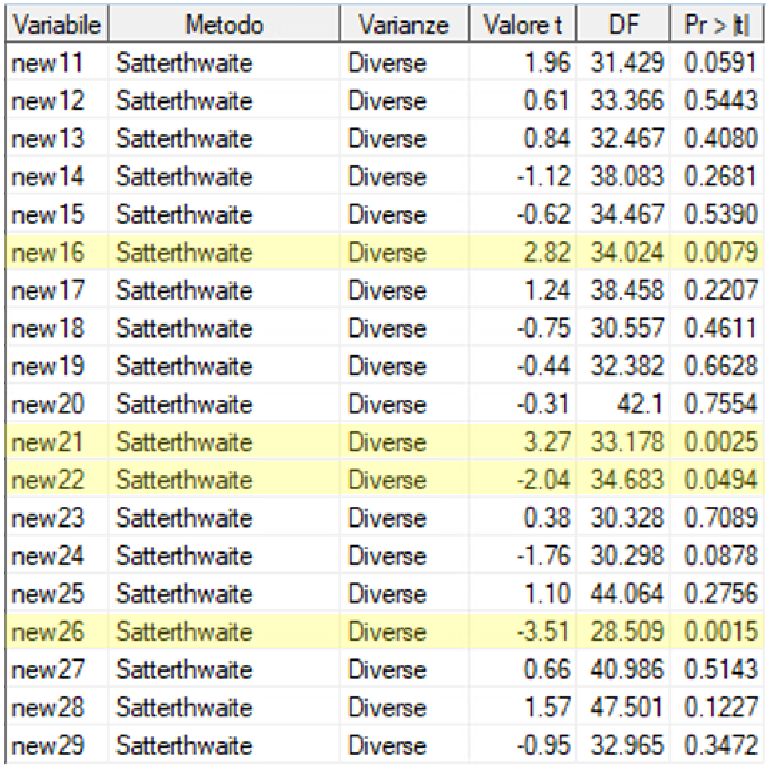
Note how variables d19, d25, d27 and d29 are common between Cluster 1 and Cluster 2. However, we notice that:

- d19 has a positive t-value in Cluster 2, while for Cluster 1 it has a negative affection;
- variables d25, d27 and d29 have negative values, while, in Cluster 1, they have positive values.

This means that variables can characterise different clusters, both accordingly or in an opposite way. Said this, we observe that the only variable with a positive value for Cluster 2 is d9, moreover, variables d11, d16 and d17 have strongly negative values.

CLUSTER 3

As previously seen, Cluster 3 has a similar size to Cluster 2. It has only 4 variables that meet the α threshold: d16, d21, d22, d26.

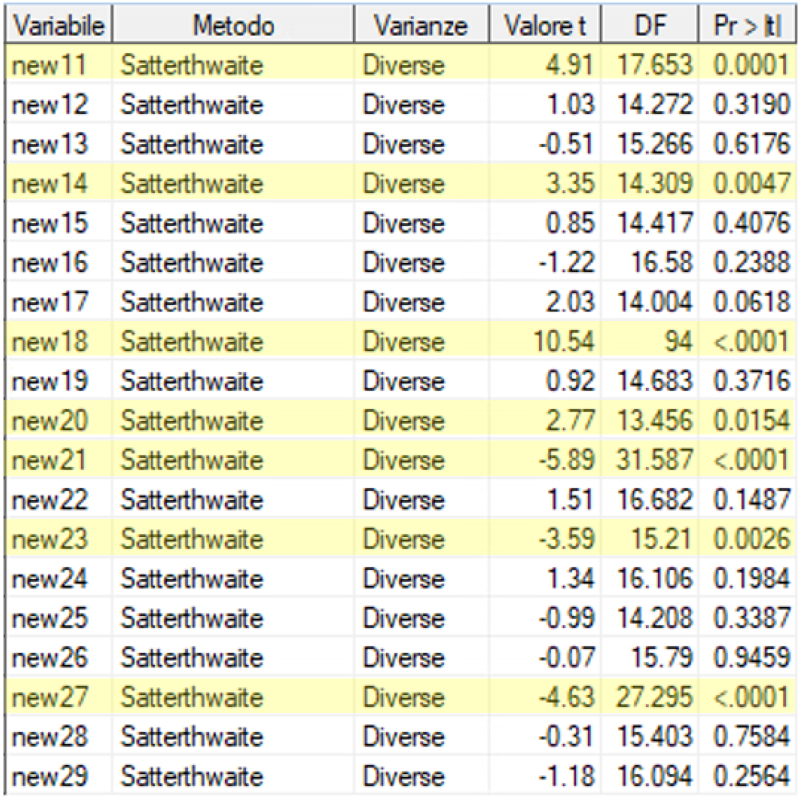


Variabile	Metodo	Varianze	Valore t	DF	Pr > |t|
new11	Satterthwaite	Diverse	1.96	31.429	0.0591
new12	Satterthwaite	Diverse	0.61	33.366	0.5443
new13	Satterthwaite	Diverse	0.84	32.467	0.4080
new14	Satterthwaite	Diverse	-1.12	38.083	0.2681
new15	Satterthwaite	Diverse	-0.62	34.467	0.5390
new16	Satterthwaite	Diverse	2.82	34.024	0.0079
new17	Satterthwaite	Diverse	1.24	38.458	0.2207
new18	Satterthwaite	Diverse	-0.75	30.557	0.4611
new19	Satterthwaite	Diverse	-0.44	32.382	0.6628
new20	Satterthwaite	Diverse	-0.31	42.1	0.7554
new21	Satterthwaite	Diverse	3.27	33.178	0.0025
new22	Satterthwaite	Diverse	-2.04	34.683	0.0494
new23	Satterthwaite	Diverse	0.38	30.328	0.7089
new24	Satterthwaite	Diverse	-1.76	30.298	0.0878
new25	Satterthwaite	Diverse	1.10	44.064	0.2756
new26	Satterthwaite	Diverse	-3.51	28.509	0.0015
new27	Satterthwaite	Diverse	0.66	40.986	0.5143
new28	Satterthwaite	Diverse	1.57	47.501	0.1227
new29	Satterthwaite	Diverse	-0.95	32.965	0.3472

The variable d16 is common to Cluster 2, while d26 is shared with Cluster 1. However, in this case, both d16 and d26 have opposite signs compared with the previous clusters. We also see that d26 is the variable with the lowest value for this cluster, while d21 is the one with the highest value.

CLUSTER 4

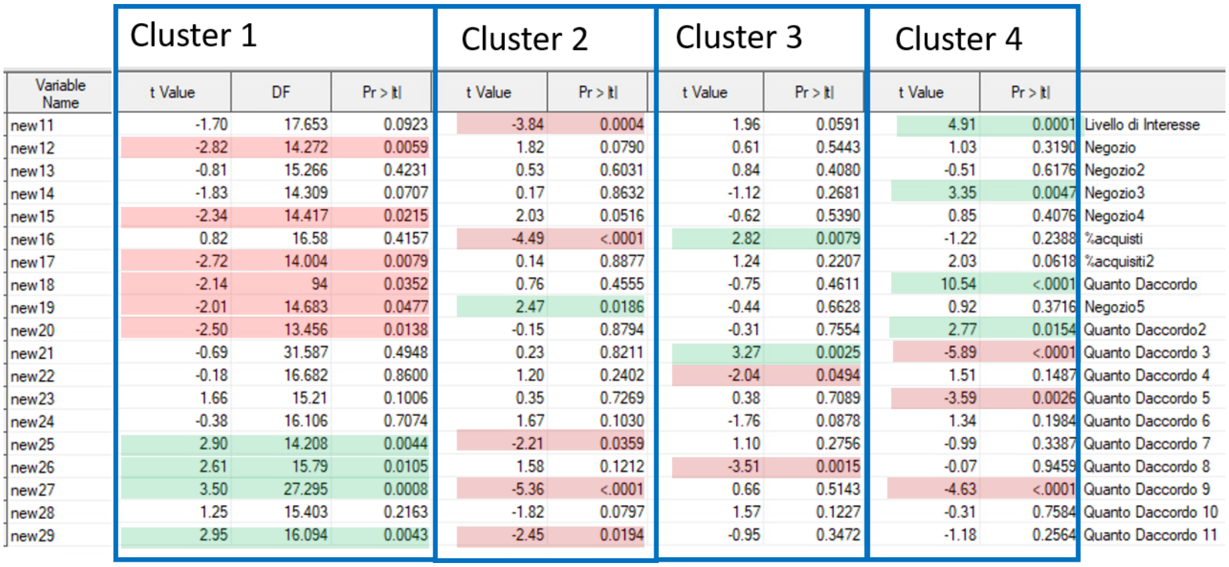
Eventually, Cluster 4, which is the smallest in terms of size, has seven variables meeting the threshold-limitation: d11, d14, d18, d20, d21, d23, d27.



Variabile	Metodo	Varianze	Valore t	DF	Pr > |t|
new11	Satterthwaite	Diverse	4.91	17.653	0.0001
new12	Satterthwaite	Diverse	1.03	14.272	0.3190
new13	Satterthwaite	Diverse	-0.51	15.266	0.6176
new14	Satterthwaite	Diverse	3.35	14.309	0.0047
new15	Satterthwaite	Diverse	0.85	14.417	0.4076
new16	Satterthwaite	Diverse	-1.22	16.58	0.2388
new17	Satterthwaite	Diverse	2.03	14.004	0.0618
new18	Satterthwaite	Diverse	10.54	94	<.0001
new19	Satterthwaite	Diverse	0.92	14.683	0.3716
new20	Satterthwaite	Diverse	2.77	13.456	0.0154
new21	Satterthwaite	Diverse	-5.89	31.587	<.0001
new22	Satterthwaite	Diverse	1.51	16.682	0.1487
new23	Satterthwaite	Diverse	-3.59	15.21	0.0026
new24	Satterthwaite	Diverse	1.34	16.106	0.1984
new25	Satterthwaite	Diverse	-0.99	14.208	0.3387
new26	Satterthwaite	Diverse	-0.07	15.79	0.9459
new27	Satterthwaite	Diverse	-4.63	27.295	<.0001
new28	Satterthwaite	Diverse	-0.31	15.403	0.7584
new29	Satterthwaite	Diverse	-1.18	16.094	0.2564

The variable d11 is common to Cluster 2, but with positive, therefore opposite, sign within it. Variables d18 and d20 are common to Cluster 1, with a positive and therefore opposite value. The variable d21 is common to Cluster 3, once again with a negative and therefore opposite sign. In the end, variable d27 is common to both Cluster 2 and Cluster 1, with a negative value. In particular, we can say there are strongly polarised t-values, for example both d11 and d18 have very high values, while d21 and d27 have heavily low values.

Summing up, the table below shows the t-Values and p-Values for all clusters, with positive values in green and negative values in red.



	Cluster 1			Cluster 2		Cluster 3		Cluster 4		
Variable Name	t Value	DF	Pr > |t|	t Value	Pr > |t|	t Value	Pr > |t|	t Value	Pr > |t|	
new11	-1.70	17.653	0.0923	-3.84	0.0004	1.96	0.0591	4.91	0.0001	Livello di Interesse
new12	-2.82	14.272	0.0059	1.82	0.0790	0.61	0.5443	1.03	0.3190	Negozio
new13	-0.81	15.266	0.4231	0.53	0.6031	0.84	0.4080	-0.51	0.6176	Negozio2
new14	-1.83	14.309	0.0707	0.17	0.8632	-1.12	0.2681	3.35	0.0047	Negozio3
new15	-2.34	14.417	0.0215	2.03	0.0516	-0.62	0.5390	0.85	0.4076	Negozio4
new16	0.82	16.58	0.4157	-4.49	<.0001	2.82	0.0079	-1.22	0.2388	%acquisti
new17	-2.72	14.004	0.0079	0.14	0.8877	1.24	0.2207	2.03	0.0618	%acquisiti2
new18	-2.14	94	0.0352	0.76	0.4555	-0.75	0.4611	10.54	<.0001	Quanto Daccordo
new19	-2.01	14.683	0.0477	2.47	0.0186	-0.44	0.6628	0.92	0.3716	Negozio5
new20	-2.50	13.456	0.0138	-0.15	0.8794	-0.31	0.7554	2.77	0.0154	Quanto Daccordo2
new21	-0.69	31.587	0.4948	0.23	0.8211	3.27	0.0025	-5.89	<.0001	Quanto Daccordo 3
new22	-0.18	16.682	0.8600	1.20	0.2402	-2.04	0.0494	1.51	0.1487	Quanto Daccordo 4
new23	1.66	15.21	0.1006	0.35	0.7269	0.38	0.7089	-3.59	0.0026	Quanto Daccordo 5
new24	-0.38	16.106	0.7074	1.67	0.1030	-1.76	0.0878	1.34	0.1984	Quanto Daccordo 6
new25	2.90	14.208	0.0044	-2.21	0.0359	1.10	0.2756	-0.99	0.3387	Quanto Daccordo 7
new26	2.61	15.79	0.0105	1.58	0.1212	-3.51	0.0015	-0.07	0.9459	Quanto Daccordo 8
new27	3.50	27.295	0.0008	-5.36	<.0001	0.66	0.5143	-4.63	<.0001	Quanto Daccordo 9
new28	1.25	15.403	0.2163	-1.82	0.0797	1.57	0.1227	-0.31	0.7584	Quanto Daccordo 10
new29	2.95	16.094	0.0043	-2.45	0.0194	-0.95	0.3472	-1.18	0.2564	Quanto Daccordo 11

4.6 Chi-square test

So far, our analysis has only involved numerical variables. However, as we have already mentioned during the description of the survey, some responses are categorical variables.

In this section, we carry out the **Chi-square test** over our categorical variables, which is useful to verify whether we can accept or reject the null hypothesis, similarly to what we did to the numerical variables with the t-test.

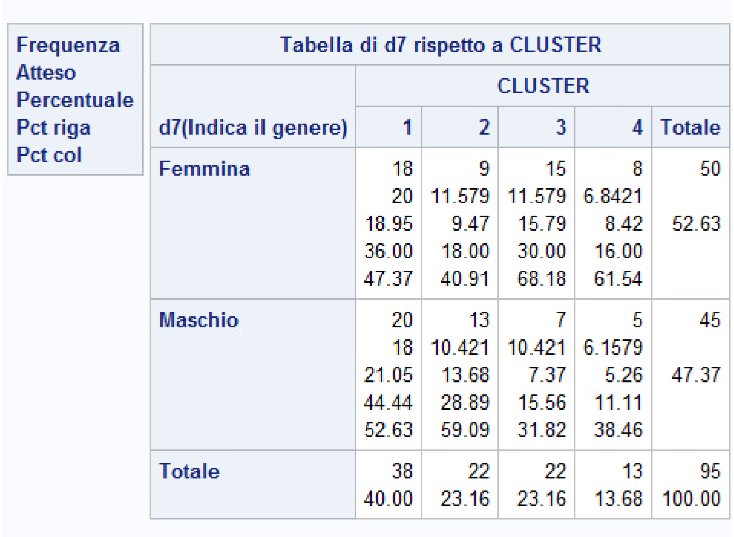
Here is a recall to the categorical variables of our survey, over which we will perform Chi-square:

- d7: Gender;
- d8: Age;
- d9: profession;
- d10: geographical location.

We already said that Chi-Square is the measure of the difference between the observed frequency and the expected frequency of a set of variables.

We performed this test for each categorical variable for each cluster.

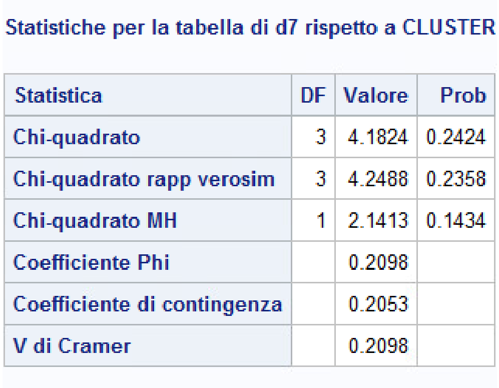
As presented in the following images, in Cluster 3, there is an over-representation of samples who answered “Female” to the Gender question. So, similarly to what we did during the t-test, we set a maximum threshold **α=0.10** for the **p-value**, in order to reject the null hypothesis for a variable.



Frequenza
Atteso
Percentuale
Pct riga
Pct col

Tabella di d7 rispetto a CLUSTER

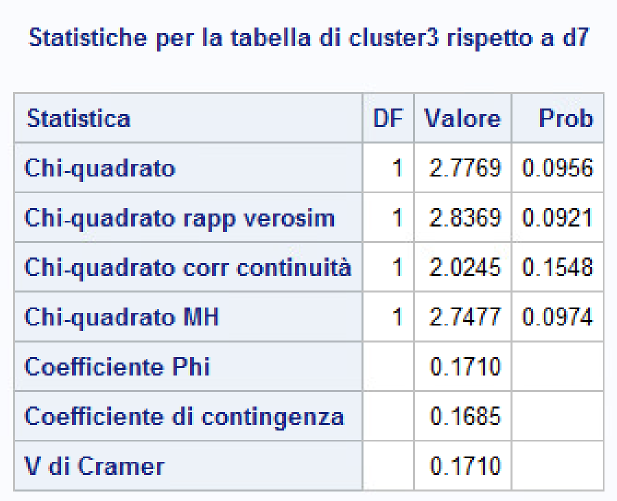
	CLUSTER				
d7(Indica il genere)	1	2	3	4	Totale
Femmina	18 20 18.95 36.00 47.37	9 11.579 9.47 18.00 40.91	15 11.579 15.79 30.00 68.18	8 6.8421 8.42 16.00 61.54	50 52.63
Maschio	20 18 21.05 44.44 52.63	13 10.421 13.68 28.89 59.09	7 10.421 7.37 15.56 31.82	5 6.1579 5.26 11.11 38.46	45 47.37
Totale	38 40.00	22 23.16	22 23.16	13 13.68	95 100.00



Statistiche per la tabella di d7 rispetto a CLUSTER

Statistica	DF	Valore	Prob
Chi-quadrato	3	4.1824	0.2424
Chi-quadrato rapp verosim	3	4.2488	0.2358
Chi-quadrato MH	1	2.1413	0.1434
Coefficiente Phi		0.2098	
Coefficiente di contingenza		0.2053	
V di Cramer		0.2098	

In the following table, we see that the result of this analysis over Cluster 3 shows a p-value of 0.0956, which meets the threshold that we set, therefore we can state that the null hypothesis is rejected. Said this, we now see the greater presence of females as a characterisation of Cluster 3.



Statistiche per la tabella di cluster3 rispetto a d7

Statistica	DF	Valore	Prob
Chi-quadrato	1	2.7769	0.0956
Chi-quadrato rapp verosim	1	2.8369	0.0921
Chi-quadrato corr continuità	1	2.0245	0.1548
Chi-quadrato MH	1	2.7477	0.0974
Coefficiente Phi		0.1710	
Coefficiente di contingenza		0.1685	
V di Cramer		0.1710	

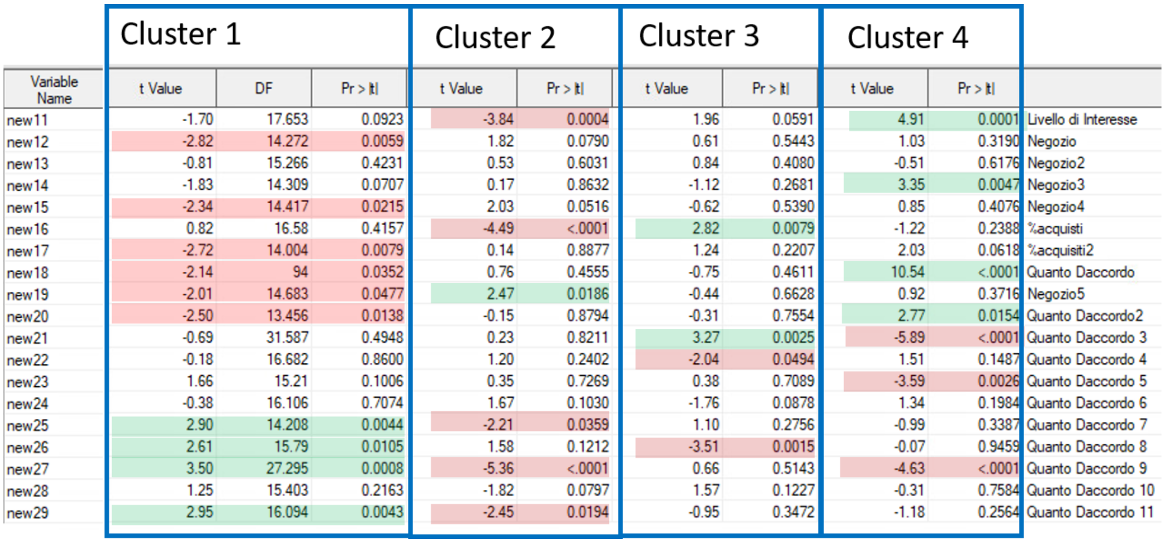
We carry out the same analysis to the other clusters as well, concerning the variable d7, but they do not meet the threshold imposition.

We also repeated this test on all the other categorical variables over the four clusters, but we did not find any categorical variable with a lower p-value than the threshold α, therefore we cannot infer any characterisation on the basis of the categorical variables, apart from what has already been said about the gender-variable d7 on Cluster 3.

This behaviour actually respects the expectations that we already described in the preliminary phase: since not having a well distributed audience over these variables, except for the gender, it is difficult to obtain statistically significant results.

5 Cluster description

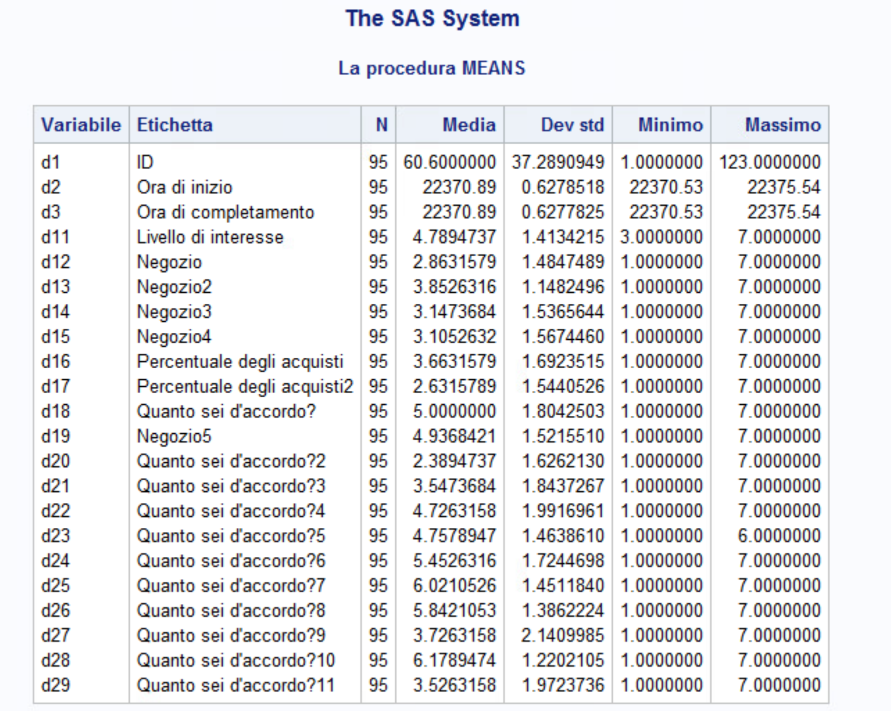
At the end of the clustering process, we found ourselves with four clusters. For each of them, we performed a ***proc ttest*** obtaining the following table:



	Cluster 1			Cluster 2		Cluster 3		Cluster 4		
Variable Name	t Value	DF	Pr > |t|	t Value	Pr > |t|	t Value	Pr > |t|	t Value	Pr > |t|	
new11	-1.70	17.653	0.0923	-3.84	0.0004	1.96	0.0591	4.91	0.0001	Livello di Interesse
new12	-2.82	14.272	0.0059	1.82	0.0790	0.61	0.5443	1.03	0.3190	Negozio
new13	-0.81	15.266	0.4231	0.53	0.6031	0.84	0.4080	-0.51	0.6176	Negozio2
new14	-1.83	14.309	0.0707	0.17	0.8632	-1.12	0.2681	3.35	0.0047	Negozio3
new15	-2.34	14.417	0.0215	2.03	0.0516	-0.62	0.5390	0.85	0.4076	Negozio4
new16	0.82	16.58	0.4157	-4.49	<.0001	2.82	0.0079	-1.22	0.2388	%acquisti
new17	-2.72	14.004	0.0079	0.14	0.8877	1.24	0.2207	2.03	0.0618	%acquisiti2
new18	-2.14	94	0.0352	0.76	0.4555	-0.75	0.4611	10.54	<.0001	Quanto Daccordo
new19	-2.01	14.683	0.0477	2.47	0.0186	-0.44	0.6628	0.92	0.3716	Negozio5
new20	-2.50	13.456	0.0138	-0.15	0.8794	-0.31	0.7554	2.77	0.0154	Quanto Daccordo2
new21	-0.69	31.587	0.4948	0.23	0.8211	3.27	0.0025	-5.89	<.0001	Quanto Daccordo 3
new22	-0.18	16.682	0.8600	1.20	0.2402	-2.04	0.0494	1.51	0.1487	Quanto Daccordo 4
new23	1.66	15.21	0.1006	0.35	0.7269	0.38	0.7089	-3.59	0.0026	Quanto Daccordo 5
new24	-0.38	16.106	0.7074	1.67	0.1030	-1.76	0.0878	1.34	0.1984	Quanto Daccordo 6
new25	2.90	14.208	0.0044	-2.21	0.0359	1.10	0.2756	-0.99	0.3387	Quanto Daccordo 7
new26	2.61	15.79	0.0105	1.58	0.1212	-3.51	0.0015	-0.07	0.9459	Quanto Daccordo 8
new27	3.50	27.295	0.0008	-5.36	<.0001	0.66	0.5143	-4.63	<.0001	Quanto Daccordo 9
new28	1.25	15.403	0.2163	-1.82	0.0797	1.57	0.1227	-0.31	0.7584	Quanto Daccordo 10
new29	2.95	16.094	0.0043	-2.45	0.0194	-0.95	0.3472	-1.18	0.2564	Quanto Daccordo 11

To decide whether a variable is meaningful to us or not, we defined a threshold over the p-Value **Pr>|t|**. We opted for a threshold of **0.05**, meaning that those variables with Pr>|t| smaller than 0.0 5 were considered as characterizing for a certain cluster and are highlighted in the table above. Then we considered if the **t-Value** of each meaningful variable is positive or negative, to understand what has been the impact of the replies to that question with regards to the clusters. In particular, meaningful variables with negative t-Value are highlighted in red, while those with positive t-Value are highlighted in green.

In order to understand these results, each t-Value was compared in relation to its variable's average over the entire dataset. The table containing the average values for the dataset's variables is re-shown below:



The SAS System

La procedura MEANS

Variabile	Etichetta	N	Media	Dev std	Minimo	Massimo
d1	ID	95	60.6000000	37.2890949	1.0000000	123.0000000
d2	Ora di inizio	95	22370.89	0.6278518	22370.53	22375.54
d3	Ora di completamento	95	22370.89	0.6277825	22370.53	22375.54
d11	Livello di interesse	95	4.7894737	1.4134215	3.0000000	7.0000000
d12	Negozio	95	2.8631579	1.4847489	1.0000000	7.0000000
d13	Negozio2	95	3.8526316	1.1482496	1.0000000	7.0000000
d14	Negozio3	95	3.1473684	1.5365644	1.0000000	7.0000000
d15	Negozio4	95	3.1052632	1.5674460	1.0000000	7.0000000
d16	Percentuale degli acquisti	95	3.6631579	1.6923515	1.0000000	7.0000000
d17	Percentuale degli acquisti2	95	2.6315789	1.5440526	1.0000000	7.0000000
d18	Quanto sei d'accordo?	95	5.0000000	1.8042503	1.0000000	7.0000000
d19	Negozio5	95	4.9368421	1.5215510	1.0000000	7.0000000
d20	Quanto sei d'accordo?2	95	2.3894737	1.6262130	1.0000000	7.0000000
d21	Quanto sei d'accordo?3	95	3.5473684	1.8437267	1.0000000	7.0000000
d22	Quanto sei d'accordo?4	95	4.7263158	1.9916961	1.0000000	7.0000000
d23	Quanto sei d'accordo?5	95	4.7578947	1.4638610	1.0000000	6.0000000
d24	Quanto sei d'accordo?6	95	5.4526316	1.7244698	1.0000000	7.0000000
d25	Quanto sei d'accordo?7	95	6.0210526	1.4511840	1.0000000	7.0000000
d26	Quanto sei d'accordo?8	95	5.8421053	1.3862224	1.0000000	7.0000000
d27	Quanto sei d'accordo?9	95	3.7263158	2.1409985	1.0000000	7.0000000
d28	Quanto sei d'accordo?10	95	6.1789474	1.2202105	1.0000000	7.0000000
d29	Quanto sei d'accordo?11	95	3.5263158	1.9723736	1.0000000	7.0000000

5.1 Cluster 1 - TRADITIONAL SHOPPERS

This cluster is characterized by ten variables and so it is the one with the widest range of meaningful parameters.

First variable to be considered (**d12**) is the one containing the responses to *"In order to buy luxury products, do you usually rely on physical shops [0] or online shops [7]?"*. In this case, t-Value is -2.82. The average value for this variable considering the whole dataset is 2.86, so this variable characterizes Cluster 1 by expressing its tendency to strongly prefer physical stores when it comes to luxury shopping.

Then, variable (**d15**) representing *"In order to buy clothes, select if you mostly purchase in physical stores [0] or online [7]"* has a negative t-Value of -2.34. The average value for this variable in the dataset is 3.11, so, according to this variable, Cluster 1 is characterized by a strong preference for physical shops for clothing shoppings in general.

The next variable (**d17**) collects the answers to the question *"Select how much of your clothes online shoppings concerns luxury brands"*. The average value for this variable is 2.63 and its t-Value within Cluster 1 is -2.72, so this cluster is characterized by people who do not buy online clothes by luxury brands, confirming the knowledge gathered by variable d12.

Another meaningful variable for Cluster 1 (**d18**) is the one about the question *"By relying on shopping online, do you have the possibility to buy from brands not easily reachable otherwise? 0-7"*. For this one, the dataset's average is 5 and t-Value is -2.14, meaning that people in this cluster do not reckon that shopping online permits them to buy from particular brands that would be unreachable otherwise.

To characterize Cluster 1, we also found the variable (**d19**) summarizing the replies to *"In order to find discounts or promotions, select if you rely mostly on physical shops [0], like Outlets, or online shops[7]"*. The average value for the dataset is 4.94. Since the t-Value of this variable is negative, -2.01, we understand that people in this cluster are more likely than average to rely on physical shops even to find promotions.

In addition, Cluster 1 is typified by some agreement-level variables, like the one to the sentence *"Buying online is the same as buying in physical stores [0-7]"* (**d20**). In this case, the average is 2.39 so, given that t-Value is -2.50, people of this cluster do not think that buying online is the same experience as buying in physical shops.

Another opinion question that is meaningful to this cluster is about *"I prefer to buy in physical shops to try on the clothes that I wish to buy [0-7]"* (**d25**). The average of this value through the dataset is 6.02 and Cluster 1's t-Value is 2.90, therefore we can assume that the responses made by this cluster's people to this question are the closest to the maximum value (7), meaning that they consider the possibility to try on clothes such a decisive factor that this is one reason why they prefer physical shops.

On reaching the last variables to deal with, we see the one concerning *"The presence of discounts affects my willingness to buy [0-7]"* (**d26**). For this variable, average is 5.84 and t-Value is 2.61, hence we can say that Cluster 1's tendency is once again close to the maximum value, namely people of this cluster are influenced by the presence of discounts.

Then, there is the variable describing the responses to *"I do not trust paying a large amount of money online [0-7]"* (**d27**). Here, we see an average value of 3.73 and a positive t-Value, the highest one for this cluster, with a value of 3.50. This means that people of this cluster do not trust paying much money online and this is another decisive factor to their preference to buy luxury items in physical shops.

Eventually, the last variable (**d29**) typifying Cluster 1 concerns the agreement-level to the sentence *"I prefer to buy in physical stores because I can get advice [0-7]"*. This variable's average is 3.53 and, within this cluster, its t-Value is 2.95. This allows the interpretation that people of this cluster consider the possibility to be advised as an additional decisive factor to prefer physical shopping.

Summing up, the sample of the dataset that constitutes Cluster 1 contains those people who do not buy luxury brands' clothes from online stores. They also do not reckon that shopping online permits them to buy from particular brands that would be unreachable otherwise; this might be because they do not need online stores because they live in cities offering the same physical ones or for other reasons. People of Cluster 1 show the tendency to strongly prefer physical stores when it comes to luxury shopping and for clothing shoppings in general. In addition, people of this cluster do not think that buying online is equivalent to doing so in physical shops and they also do not trust paying much money online. In particular, according to the t-Values, this last factor seems to be the strongest one to explain their preference for physical shops. Other factors are the possibility to try on clothes and the possibility to be advised. On the basis of these aspects, we decided to name Cluster 1 "**Traditional Shoppers**", because we felt like people belonging to this cluster have usual behaviour and thoughts regarding shopping habits. Moreover, people of this cluster are influenced by the presence of discounts and promotions and they are more likely to bring on their promotions' hunt relying on physical shops.

5.2 Cluster 2 - ONLINE SAVERS

For what concerning Cluster 2, as highlighted in the table above, the variables that describe the most this segment are six.

The first one (**d11**) is the variable representing the following question of our questionnaire: *"How much are you interested in the luxury market concerning the*

Fashion Industry? [0-7]". It is important to specify that this variable is our screening question and we set a threshold (3>) over it, in order to capture our intended audience, consequently it is the only one that has been rescaled. In order to proceed with the cluster description, the average value for this variable over the answers gained is 4.79 and the t-Value referred to this specific cluster is -3.84. The meaning of these values is that this cluster contains those people who showed the lowest interest in the luxury market.

The second one (**d16**) is represented by the question *"Select how much of your clothing shoppings in physical stores concerns luxury brands: 0-7"*. In this case the average response per question is 3.66 and the t-Value for the cluster is negative, -4.49, meaning that people of this cluster do not use to buy luxury clothes in physical shops. It is interesting to highlight that the question *"Select how much of your clothes online shoppings concerns luxury brands: 0-7"*, that would have been useful to specifically describe this cluster together with variable d16, does not characterize it.

The third variable (**d19**) is the representation of the question *"In order to find discounts or promotions, select if you rely mostly on physical shops [0], like Outlets, or online shops [7]"*. Here, the average response is 4.94 and its t-Value for this cluster is 2.47, meaning that people of this cluster look for discounts online.

Another one (**d25**) is the agreement question *"I prefer to buy in physical shops to try on the clothes that I wish to buy [0-7]"*, where the average response is 6.02 and the t-Value for the cluster is negative, -2.21. We interpret these values by saying people of this cluster are not interested in trying on the clothes before buying them.

For this cluster, we also have the variable asking for agreement-level to the statement *"I do not trust paying a large amount of money online [0-7]"* (**d25**). In this case, the average response is 3.73 and the t-Value is -5.36, meaning that people in this cluster are not afraid of spending money online.

The last variable (**d29**) summarizes the responses to *"I prefer to buy in physical stores because I can get advice [0-7]"*. The average response is 3.53 and the t-Value for the cluster is -2.45, which means that this cluster is not interested in getting advised.

In summary, this cluster contains people who are the least interested in the luxury market, neither in physical nor online shops. We labeled them "**Online Savers**" because they look for discounts online and use to buy clothes or accessories online, not being afraid of spending money on websites. Moreover, they are not interested in getting personal advice and, for them, it is not important to try on clothes in order to buy them, and these are additional reasons not to prefer physical shops.

5.3 Cluster 3 - LUXURY BUYERS

Earlier, we found out that this cluster is affected by an unbalanced situation according to the gender distribution, because 68% of its people are females, and we also realised that Cluster 3's characterisation depends on 4 variables.

The first variable (**d16**) is the representation of the question *"Select how much of your clothing shoppings in physical stores concerns luxury brands: [0-7]"*. The general average response is 3.66 and its t-Value for this cluster is positive: 2.82. This means that people belonging to this cluster, when buying clothes in physical shops, tend to buy luxury items.

Another one (**d21**) concerns the agreement-question *"When I buy online, I only buy from the brand's official website [0-7]"*. The general average is 3.54 and the t-Value is 3.27: people of this cluster, when buying online, tend to buy from official websites.

The third variable (**d22**) summarizes the responses to *"I prefer to buy online because it saves me time [0-7]"*. The average value for this variable is 4.73 and the t-Value with Cluster 3 is negative -2.04, meaning that people of this cluster do not prefer to buy online in order to save time.

The last variable (**d26**) is about agreement level to the sentence: *"The presence of discounts affects my willingness to buy [0-7]"*. The general average is 5.84 and the t-Value is negative, -3.51. We interpreted this as people in this cluster not being influenced by the presence of discounts or promotions.

All in all, this cluster contains those people who often buy luxury clothing when they go shopping. Moreover, when they buy online, they only buy from the official brand sites. We consider this behaviour as typical of those buyers who are used to luxury shopping, because they are looking for the shopping experience in addition to products of the highest quality, quality that they want to see and touch, even if this means spending more money or longer time. We cannot say for sure if these people do not like to buy online, however we clearly understand that, concerning luxury shopping, they trust only official vendors, probably because they are an additional safety to the authenticity of the items and to their being current-season. In fact, new, seasonal pieces tend to be available online on the brand's official websites firstly and, only later, they are accessible at resellers' websites. This could be a sign of how much people in this cluster pay attention to the exclusivity factor. For these reasons, we decided to call this cluster "**Luxury Buyers**", as we think this cluster contains those people who are used to luxury shopping.

5.4 Cluster 4 - ONLINE HUNTERS

The characterisation of this cluster depends on 7 variables.

The first variable (**d11**) concerns the question *"How much are you interested in the luxury market concerning the Fashion Industry? [0-7]"*. The general average is 4.79 and the t-Value is 4.91, which means that people of this cluster show a strong interest in the luxury market.

The second variable (**d14**) is about *"In order to buy accessories (bags, belts, shoes, etc.), select if you mostly purchase in physical stores [0] or online [7]"*. The average value is 3.15 and the t-Value is 3.35. This is interpreted as people of this cluster preferring to buy online when looking for accessories.

Another variable (**d18**) summarizes the responses to *"By relying on shopping online, do you have the possibility to buy from brands not easily reachable otherwise? [0-7]"*. In this case, average is 5 and the t-Value within this cluster is strongly positive: 10.54. Thanks to this feature, we realised that people of this cluster feel that online shopping is crucial to reach and buy what they look for.

The fourth one (**d20**) describes agreement-level to the statement: *"Buying online is the same as buying in physical stores [0-7]"*. The general average is 2.39 and the t-Value is 2.77, meaning that people of this cluster consider buying in physical stores not so different to buying online, compared to the answers given by the rest of the dataset.

The next variable (**d21**) concerns agreement-question *"When I buy online, I only buy from the brand's official website [0-7]"*. The average value is 3.54 and its t-Value is strongly negative: -5.89. This can be read as people of this cluster not at all buying exclusively from brands' official websites.

We also have to characterise Cluster 4 the variable about agreement to "Buying in a physical store allows me to support the community I live in [0-7]" (**d23**). This has an average of 4.76 and t-Value of -3.59, meaning that people of this cluster do not feel that buying in physical shops helps their community.

The last variable (**d27**) is about *"I do not trust paying a large amount of money online [0-7]"*. Here, the general average is 3.73 and its t-Value is -4.63, so we can say that people of this cluster have no problem with spending large amounts of money online.

Summarizing, this cluster contains people who are particularly interested in luxury shopping, especially concerning accessories markets. A key characteristic of this cluster is the need to use online shopping to carry out its people's purchases, indeed it is mainly through the online market that they are able to buy their wanted items, that they would not be able to buy otherwise. The reasons for this could be many, but the most likely is that the goods in which this cluster is interested are sold almost exclusively online, such as the market for luxury sneakers and shoe-trainers. in particular, we know that this market has no official channels, neither in shops nor online, and the most sought-after models are often only sold through online auctions. Moreover, this cluster contains people who spend large amounts of money online, mainly on retailers' websites, often without even considering the physical shop as a social element for their community, probably because the desire to own a certain object, limited and exclusive, is much stronger than other considerations. For these reasons we named this cluster "**Online Hunters**".

Conclusion

As previously seen in the Clustering section, the 40% of our survey's audience lies in Cluster 1, making it the largest cluster. Then, Cluster 2 and Cluster 3 have the same size, with 23% of the respondents, while Cluster 4 is the smallest one, with 13% of the samples.

The main goal of our research was to understand whether the part of the audience that is interested in luxury shopping prefers to buy online or in physical shops, plus, we aimed at understanding and designing targeted strategies.

Since *Traditional Shoppers Cluster* is the most populated, we reckon that maintaining physical shops as a strong channel distribution is the best choice.

Concerning the online channels, we thought of different strategies accordingly with the characteristics of the other clusters. The *Luxury Buyers Cluster* highlights that the main online channel for luxury clothes shopping is the one of official websites, while, for accessories, non-official stores and retailers might be a good solution, according to what *Online Hunters Cluster* suggests, in particular for limited-editions and brand-collab.

Another possible strategy to reach the *Online Savers Cluster* could be to dispatch end-of-series or earlier season's collections through non-official retailers as discounted items.

Bibliography

1. https://www.italiaonline.it/risorse/settore-lusso-ecco-come-ripartira-grazie-all-e-commerce-2382
2. https://www.mglobale.it/analisi-di-mercato/tutte-le-news/osservatorio-altagamma-lusso-in-calo-del-20-.kl
3. https://medium.com/@dareyadewumi650/understanding-the-role-of-eigenvectors-and-eigenvalues-in-pca-dimensionality-reduction-10186dad0c5c
4. https://stats.stackexchange.com/questions/93905/can-averaging-all-the-variables-be-seen-as-a-crude-form-of-pca
5. https://math.stackexchange.com/questions/2473219/positive-eigenvectors-for-nonnegative-matrices
6. https://online.stat.psu.edu/stat505/lesson/11
7. https://en.wikipedia.org/wiki/Likert_scale