BROWSING SYSTEMS

Customer Behaviour Analysis

Group Project
Course // 81613 Business Intelligence
Professor // Furio Camillo
Students //

Erik Arutyunyan
Laura Canali
Gabriele Politi
Leonardo Vincenzi

Table of Contents

1. Introduction 3

1.1 Survey 4
1.2 Questionnaire 4

2. Preliminary Analysis 7

2.1 Creation of the library 7
2.2 Importation of the dataset 7
2.3 Labelling process 8
2.4 Proc Contents 9
2.5 Proc Means 10
2.6 Proc Freq 12

3. Principal Component Analysis 19

3.1 Elimination of Size Effect 22
3.2 Final PCA 24

4. Clustering 27

4.1 Clustering method and formation 27
4.2 Preparation for the T-test 29
4.3 T-test 30
4.4 Chi-square test 32

5. Cluster description 36

5.1 General analysis 36
5.2 Cluster 1 37
5.3 Cluster 2 37
5.4 Cluster 3 38
5.5 Cluster 4 38
5.6 Cluster 5 39

Conclusion 40

1. Introduction

In the past decades, the technological innovation caused great changes in humans' life. One of the greatest is the internet invention, which allowed people to fast communication, discovering and exchange of new information. Billions of people all over the world use internet and, to do so, they need to go through a browser.

The aim of this paper is to understand and analyze the preferences of potential consumers concerning browsers and their features.

The motivation that lays as the basis of this study derives from the existence of many different browsers people can choose between. It has been noticed, talking with people from different countries and backgrounds, that often there is little to no consciousness about how many options are available. There are in fact many factors influencing the options to choose between. i.e., tech companies usually give people a pre-installed browser depending on the devices' brand or on the location where they have been bought. Even though there are some exceptions, for example in countries where certain browsers are not allowed, people still have the chance to pick their favorite between a wide variety of possibilities, in which, as it is easy to guess, the most popular ones are Google Chrome, Safari, Internet Explorer, and Firefox.

This study will therefore go on to identify the elements that impact the preferences and choices of potential or current consumers, dividing them into subgroups in order to better identify the features that need to be enhanced to meet the needs of each of them.

1.1 Survey

This paper's analysis is based on the data collected through the administration of a questionnaire via Qualtrics.

The dataset obtained is formed by a total of 100 respondents. To better describe their distribution, some data were collected about age, gender and provenience. More specifically, 60% of the respondents is in the age gap 18-24, 33% is in the 25-34 one, 4% is 35-44, and 3% is 45-60. About provenience, 53% is from Italy, 6% is from the region called Intra-EU, and 41% is from the region Extra-EU. Finally, genders are divided as follows: 47% identifying as males, 52% as females, and 1% as others.

The survey is composed by the following types of questions: 7 screening questions, an attribute importance section, and 5 socio-demographic questions.

1.2 Questionnaire

Screening Questions

1. How often do you use browsers?
    - Always
    - Often
    - Sometimes
    - Hardly ever
    - Never
2. How much time do you spend on average on browsers per day?
    - More than 6 hours
    - 4-6 hours
    - 2-4 hours
    - 1-2 hours

    - Less than 1 hour
3. What is your main reason for using browsers?
    - Personal
    - Education
    - Professional
4. Which browser do you mostly use? (single choice)
    - Google Chrome
    - Safari
    - Microsoft Edge
    - Mozilla Firefox
    - Opera
    - Yandex Browser
    - Baidu 百度
5. Do you use a browser which is already preinstalled on your device?
    - Yes, I use a preinstalled browser
    - Yes, but I also use other browsers
    - No, I prefer other browsers

Attribute Importance

Based on your experience, how important are the following features of a browser to you? How important is … (range: 1-7)

- Privacy of your data
- Security of a browser
- Feature integration (VPN, fingerprint, plugins, protection reports, etc.)
- Research reliability
- User interface (user-friendliness)
- Interface customization
- Integration with other devices (compatibility)

- Mobile application interface
- Speed of a search

Socio-Demographic Questions

1. Gender
    - Male
    - Female
    - Other
    - Prefer not to say
2. Age
    - Under 18
    - 18-25
    - 25-35
    - 35-45
    - 45-60
    - Above 60
3. What is your occupation?
    - Student (employed)
    - Student (unemployed)
    - Employed
    - Self-employed
    - Retired
    - Unemployed
4. What is your current degree of education?
    - Middle school
    - High school or college
    - Bachelor’s degree (BA)
    - Master’s degree (MA, MBA)

- PhD and higher

5. Where do you come from?
    - Italy
    - Intra-EU
    - Extra-EU

2. Preliminary Analysis

2.1 Creation of the library

Firstly, a permanent library has been created with the next settings to allow SAS to store all datasets there:

```
libname project 'C:\Users\erik.arutyunyan\Desktop\Browsers';
run;
```

The name of the permanent library is project and the directory of the created folder on the desktop is 'C:\Users\erik.arutyunyan\Desktop\Browsers'.

2.2 Importation of the dataset

Secondly, the primary dataset has been imported into SAS. The dataset has been exported from Qualtrics in .xlsx format as an Excel file. The next algorithm has been performed to import the dataset:

File → Import Data → Workbook → Browse (type: All files) → Library: work / Member: browsers → Finish

The next **data step** procedure has been performed to store the dataset in the permanent library and in the previously created folder:

```
data project.browsers;
set browsers;
run;
```

Where ‘browsers’ is the name of the previously imported dataset.

2.3 Labelling process

The dataset is composed of 19 variables (20, including ID). Special code names have been assigned to all variables. Therefore:

- 5 screening questions - qualitative variables (s1-s5)
- 9 attribute questions - quantitative variables (a1-a9)
- 5 socio-demographic questions - qualitative variables (d1-d5)

Also, a numerical identifier in a new column has been assigned to each response keeping the sequential order (ID). The next **data step** procedure has been performed to assign a label to each variable/column.

```
data project.browsers_1;
set project.browsers;
label s1='frequency';
label s2='time';
label s3='reason';
label s4='browser';
label s5='preinstalled';
label a1='imp_privacy';
label a2='imp_security';
label a3='imp_feature';
label a4='imp_search';
label a5='imp_interface';
label a6='imp_customization';
label a7='imp_integration';
label a8='imp_mobileInterface';
label a9='imp_speed';
label d1='gender';
label d2='age';
```

```
label d3='region';
label d4='occupation';
label d5='degree';
label id='id';
run;
```

Below the dataset before and after running the above-described procedure is presented:

Before: **data** project.browsers

	id	s1	s2	s3	s4	s5	a1	a2	a3	a4	a5	a6	a7
1	1	Always	4-6 hours	Education	Safari	Yes, I use a preinstalled web browser	7	5	6	7	4	4	5
2	2	Always	2-4 hours	Education	Safari	Yes, I use a preinstalled web browser	7	6	4	6	6	4	5
3	3	Always	2-4 hours	Education	Google Chrome	No, I prefer other web browsers	7	7	6	7	4	2	5
4	4	Always	2-4 hours	Education	Google Chrome	Yes, I use a preinstalled web browser	7	6	5	7	4	3	4
5	5	Always	2-4 hours	Education	Google Chrome	Yes, I use a preinstalled web browser	5	7	4	7	7	5	6

After: **data** project.browsers_1

	id	frequency	time	reason	browser	preinstalled	imp_privacy	imp_security	imp_feature	imp_search	imp_interface	imp_customization	imp_integration
1	1	Always	4-6 hours	Education	Safari	Yes, I use a preinstalled web browser	7	5	6	7	4	4	5
2	2	Always	2-4 hours	Education	Safari	Yes, I use a preinstalled web browser	7	6	4	6	6	4	5
3	3	Always	2-4 hours	Education	Google Chrome	No, I prefer other web browsers	7	7	6	7	4	2	5
4	4	Always	2-4 hours	Education	Google Chrome	Yes, I use a preinstalled web browser	7	6	5	7	4	3	4
5	5	Always	2-4 hours	Education	Google Chrome	Yes, I use a preinstalled web browser	5	7	4	7	7	5	6

It should be mentioned that the Cleaning procedure is skipped in SAS since the dataset has been cleaned directly in the .xlsx format Excel file after exporting it from Qualtrics. As a result, only not fully submitted responses have been deleted.

2.4 Proc Contents

To confirm the number of numeric and character variables and observations in the dataset the next **proc contents** procedure has been implemented:

```
proc contents data=project.browsers_1;
run;
```

As a result, 100 observations and 20 variables (9 numeric and 11 character) have been detected:

The SAS System

The CONTENTS Procedure

Data Set Name	PROJECT.BROWSERS_1	Observations	100
Member Type	DATA	Variables	20
Engine	V9	Indexes	0
Created	19/08/2023 11:03:48	Observation Length	240
Last Modified	19/08/2023 11:03:48	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
7	a1	Num	8			imp_privacy
8	a2	Num	8			imp_security
9	a3	Num	8			imp_feature
10	a4	Num	8			imp_search
11	a5	Num	8			imp_interface
12	a6	Num	8			imp_customization
13	a7	Num	8			imp_integration
14	a8	Num	8			imp_mobileInterface
15	a9	Num	8			imp_speed
16	d1	Char	6	$6.	$6.	gender
17	d2	Char	7	$7.	$7.	age
18	d3	Char	8	$8.	$8.	region
19	d4	Char	20	$20.	$20.	occupation
20	d5	Char	25	$25.	$25.	degree
1	id	Num	8			id
2	s1	Char	9	$9.	$9.	frequency
3	s2	Char	17	$17.	$17.	time
4	s3	Char	12	$12.	$12.	reason
5	s4	Char	15	$15.	$15.	browser
6	s5	Char	38	$38.	$38.	preinstalled

Moreover, **proc print** procedure can be used to get a full print version of the dataset stored in the library:

```
proc print data=project.browsers_1;
run;
```

2.5 Proc Means

Firstly, to study the descriptive statistics in terms of the quantitative variables the next **proc means** procedure has been performed using the numeric variables:

```
proc means data=project.browsers_1;
var a:;
run;
```

As a result, SAS by default displays the minimum, maximum, average and standard deviations:

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
a1	imp_privacy	100	5.7000000	1.5986105	1.0000000	7.0000000
a2	imp_security	100	5.9000000	1.4459976	1.0000000	7.0000000
a3	imp_feature	100	4.3800000	1.8022433	1.0000000	7.0000000
a4	imp_search	100	6.0500000	1.0384040	4.0000000	7.0000000
a5	imp_interface	100	5.4100000	1.3491112	2.0000000	7.0000000
a6	imp_customization	100	3.9300000	1.5586513	1.0000000	7.0000000
a7	imp_integration	100	5.0800000	1.4261625	2.0000000	7.0000000
a8	imp_mobileInterface	100	5.3900000	1.4832056	2.0000000	7.0000000
a9	imp_speed	100	6.2700000	1.2621530	1.0000000	7.0000000

Observing the mean values, SAS computes the average of the scores given by the respondents on a scale from 1 (Not Important at all) to 7 (Extremely Important), for each dimension:

1. *imp_customization* is the dimension with the lowest mean (3.93) suggesting that this dimension is the least important. *imp_customization is followed by imp_feature* (4.38) and *imp_integration* (5.08). These are the dimensions which do not play a key role in the selection of browsing services.
2. *imp_speed* (6.27) and *imp_search* (6.05) are displaying the highest mean values and therefore they represent the most important attributes. Furthermore, *imp_search* has the lowest standard deviation (1.03), and this means that as regards this dimension, respondents give similar judgments in terms of importance.
3. The maximum value attached to each dimension is 7 and the lowest value is 1 apart from *imp_interface*, *imp_integration*, *imp_mobileInterface* (2) and *imp_search* (4). In other words, none of these dimensions has been considered not important at all by the respondents and this is especially true for *imp_search*.

An alternative procedure to the **proc means** is the **proc summary**. The output of these procedures is the same, however, while the results of the **proc means** are

displayed in the result viewers, with the **proc summary** it is required to specify whether or not to save the results in a new dataset through the use of the parameter output out.

2.6 Proc Freq

Secondly, to study the descriptive statistics in terms of the qualitative variables the next **proc freq** procedure has been performed using the character variables:

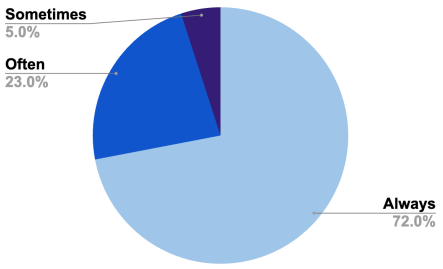
```
proc freq data=project.browsers_1;
table s:;
table d:;
run;
```

As a result, SAS by default displays the frequency and percentage of respondents, cumulative frequency and cumulative percentage:

Frequency (s1): How often do you use browsers?

According to the results, the absolute majority of respondents *always* use browsers (72%), while others (23%) consider using browsers often, and only the minority (5%) uses browsers *sometimes*. The popularity of the first answer is met by the considerably vital role of browsing systems in the contemporary world. On average, all school and university students and workers are obliged to use the systems to be able to perform their daily activities.

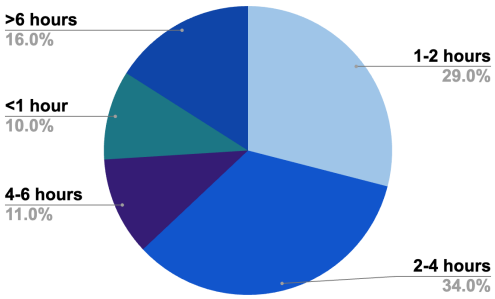
frequency				
s1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Always	72	72.00	72	72.00
Often	23	23.00	95	95.00
Sometimes	5	5.00	100	100.00



Time (s2): How much time do you spend on average on browsers per day?

The analysis shows that the majority of respondents (34%) spend on average *2-4 hours* per day on browsers, but it should be mentioned that the second most popular answer is really near to the first group, declaring that the respondents (29%) spend on average *1-2 hours* per day on browsing. The third group (16%), which is twice less than the first one, states that they spend *more than 6 hours* per day on browsers. The last 2 groups have similar results of 11% and 10%, stating that they spend on average *4-6 hours* and *less than 1 hour* per day respectively on browsing.

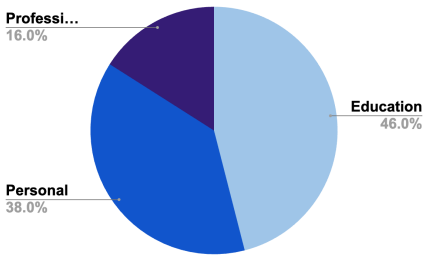
time				
s2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1-2 hours	29	29.00	29	29.00
2-4 hours	34	34.00	63	63.00
4-6 hours	11	11.00	74	74.00
Less than 1 hour	10	10.00	84	84.00
More than 6 hours	16	16.00	100	100.00



Reason (s3): What is your main reason for using browsers?

The absolute majority of respondents (46%) claim that they use browsers mainly for *Educational* purposes. The second most popular answer (38%) shows that the respondents use browsers for *Personal* reasons. An interesting fact is that the absolute minority (16%) uses browsers for *Professional* reasons. Nonetheless it had to be one of the most popular answers, it can be connected with the fact that the majority of respondents are students and young professionals.

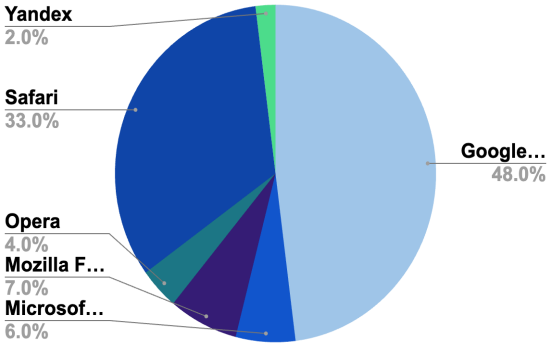
reason				
s3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Education	46	46.00	46	46.00
Personal	38	38.00	84	84.00
Professional	16	16.00	100	100.00



Browser (s4): Which browser do you mostly use?

A large variety of browsers has been chosen. The absolute majority of respondents (48%) have chosen *Google Chrome* as their daily browser, followed by *Safari* (33%). The rest of the browsers have really low rankings in the survey: *Mozilla Firefox* (7%), *Microsoft Edge* (6%), *Opera* (4%), *Yandex Browser* (2%). These results are connected with the fact that the first 2 browsers are the most popular all around the world and represent not just browsing systems, but an entire user environment with multiple functions and possibilities.

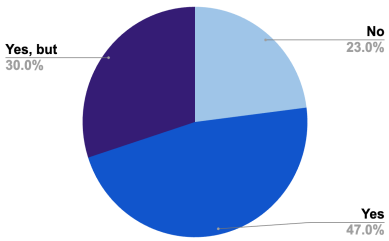
browser				
s4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Google Chrome	48	48.00	48	48.00
Microsoft Edge	6	6.00	54	54.00
Mozilla Firefox	7	7.00	61	61.00
Opera	4	4.00	65	65.00
Safari	33	33.00	98	98.00
Yandex Browser	2	2.00	100	100.00



Preinstalled (s5): Do you use a browser which is already preinstalled on your device?

The majority of respondents (47%) states that they *use a preinstalled web browser*, while the other group (30%) claims that they *use a preinstalled web browser, but they also use other browsers*. It is confirmed by the fact that for some features and opportunities, professional and personal, users usually perform their activities with the help of several browsing systems to increase their efficiency. The minority (23%) states that they *prefer other web browsers*, which can also be explained with the fact that they change preinstalled browsers since they need specific integrations and features to perform their activities.

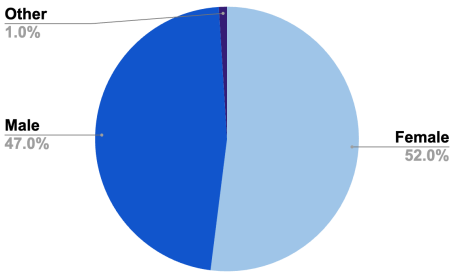
preinstalled				
s5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No, I prefer other web browsers	23	23.00	23	23.00
Yes, I use a preinstalled web browser	47	47.00	70	70.00
Yes, but I also use other web browsers	30	30.00	100	100.00



Gender (d1): How do you identify yourself?

According to the results, the gender division in this questionnaire is nearly equal, since the *Female* group (52%) is leading the *Male* group (47%) only by 5%. It also should be mentioned that 1% of the respondents chose the *Other* option in the questionnaire.

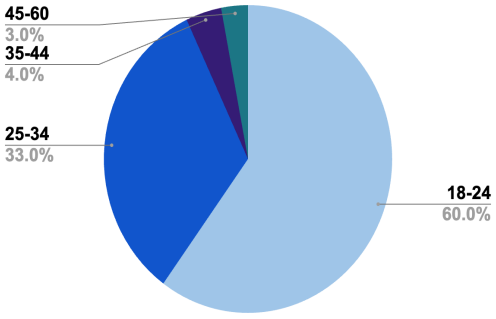
gender				
d1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	52	52.00	52	52.00
Male	47	47.00	99	99.00
Other	1	1.00	100	100.00



Age (d2): How old are you?

The table below shows that the absolute majority of respondents are aged *18-24* years old (60%), followed by the age group of *25-34* years old (33%). These results confirm the statements above, since the age group has a direct impact on specific choices and answers to other questions, including the one about the reason for using browsers. The minority age groups in this questionnaire are *35-44* (4%) and *45-60* (3%) years old.

age				
d2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
18 - 24	60	60.00	60	60.00
25 - 34	33	33.00	93	93.00
35 - 44	4	4.00	97	97.00
45 - 60	3	3.00	100	100.00

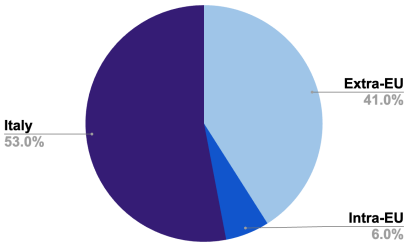


Region (d3): Where do you come from?

The results show that the majority of respondents come from *Italy* (53%), whereas the second most popular group is *Extra-EU* (41%) represented by the respondents

mainly from Armenia and Russia and some other Eastern countries. The smallest regional group is *Intra-EU* (6%) represented only by several European countries. It was decided to exclude Italy from the Intra-EU group since it was represented by a large number of respondents, and therefore it could affect the regional opinion representation of the Intra-EU group.

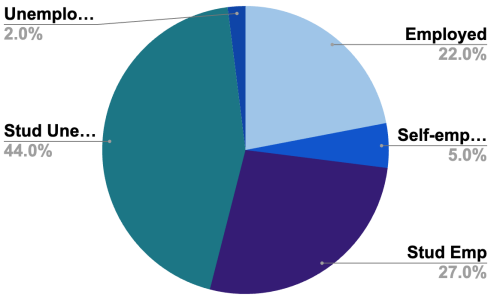
region				
d3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Extra-EU	41	41.00	41	41.00
Intra-EU	6	6.00	47	47.00
Italy	53	53.00	100	100.00



Occupation (d4): What is your occupation?

The table below shows that the majority of respondents are *Unemployed Students* (44%). It confirms the results of the question connected with the Reason for using browsers, where the majority of respondents have chosen Education as the main reason. The employed groups follow the majority opinion with nearly the same results: *Employed Students* (27%) and *Employed* (22%). The smallest groups are represented by the *Self-Employed* (5%) and *Unemployed* (2%) respondents.

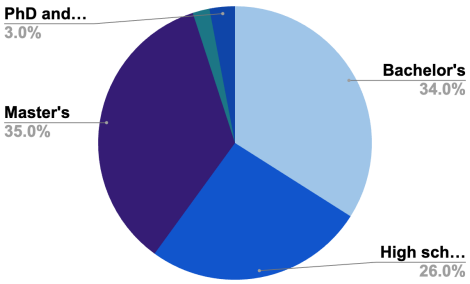
occupation				
d4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Employed	22	22.00	22	22.00
Self-employed	5	5.00	27	27.00
Student (employed)	27	27.00	54	54.00
Student (unemployed)	44	44.00	98	98.00
Unemployed	2	2.00	100	100.00



Degree (d5): What is your current degree of education?

According to the results of the procedure, the absolute majority of respondents have obtained a university diploma, totaling 69%, where *Bachelor's degree* (34%) and *Master's degree* (35%). It is followed by *High school or college* degrees (26%). Again, it confirms the results of the question connected with the Reason for using browsers, where the majority of respondents have chosen Education as the main reason. The minority groups are represented by *PhD and higher* degrees (3%) and *Middle school* degrees (2%).

degree				
d5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bachelor's degree (BA)	34	34.00	34	34.00
High school or college	26	26.00	60	60.00
Master's degree (MA, MBA)	35	35.00	95	95.00
Middle school	2	2.00	97	97.00
PhD and higher	3	3.00	100	100.00



3. Principal Component Analysis

Principal component analysis (PCA) is a statistical technique used for reducing the dimensions of large datasets and minimizing the number of variables in smaller datasets. This statistical tool is primarily used to identify a smaller number of variables, called principal components, from a larger dataset. These new variables are calculated as linear combinations of the original variables and are referred to as factor scores.

The observations' values of the new variables are seen as projections into the new principal components. Two key elements of principal analysis are *eigenvectors* and *eigenvalues*. An eigenvector represents a direction in the original space, while its corresponding. Eigenvalue represents the amount of variance explained by that eigenvector. The eigenvalues related to a component are the sum of the component's squared factor scores.

PCA is primarily used to decrease the number of highly intercorrelated variables by projecting each data point into the first principal component. This results in low-dimensional data while still maintaining the data's variation and information accuracy. To achieve this, it is crucial to first center the variables to a zero mean.

This simplifies the representations, as the center of the population's cloud coincides with the origin of the axes in the subspace.

To perform principal component analysis, the **proc princomp** command is used to summarize data and identify linear relationships.

```
proc princomp data=rest.questionario;
var a1-a9;
run;
```

```
proc princomp data=rest.questionario;
out=rest.data_coord;
var a1-a9;
run;
```

Correlation Matrix										
		a1	a2	a3	a4	a5	a6	a7	a8	a9
a1	imp_Privacy	1.0000	0.7210	0.5168	0.1856	0.0342	0.0847	0.1834	0.2628	0.3409
a2	imp_Security	0.7210	1.0000	0.4643	0.2859	0.2076	0.1851	0.3223	0.2774	0.4134
a3	imp_Feature	0.5168	0.4643	1.0000	0.1894	0.1638	0.3548	0.1335	0.2350	0.3274
a4	imp_Search	0.1856	0.2859	0.1894	1.0000	0.2232	0.1020	0.1473	0.2692	0.1206
a5	imp_Interface	0.0342	0.2076	0.1638	0.2232	1.0000	0.2444	0.2978	0.4695	0.2487
a6	imp_Customization	0.0847	0.1851	0.3548	0.1020	0.2444	1.0000	0.2706	0.1386	0.1073
a7	imp_Integration	0.1834	0.3223	0.1335	0.1473	0.2978	0.2706	1.0000	0.4101	0.0945
a8	imp_MobileInterface	0.2628	0.2774	0.2350	0.2692	0.4695	0.1386	0.4101	1.0000	0.3910
a9	imp_Speed	0.3409	0.4134	0.3274	0.1206	0.2487	0.1073	0.0945	0.3910	1.0000

The highest correlation is between the variables “imp_Security” and “imp_Privacy”. It is a meaningful result because it may represent the customers prefer using browsers that can give a high value of privacy and security. The lowest value is the correlation between the variables “imp_Interface” and “imp_Privacy”.

Consequently, we analyze the *Eigenvalues of the Correlation Matrix* to assess the number of useful components to interpret our data. Our variables are 9, so we have a table composed of 9 principal components. PCA, when you have a k-dimensional data, generates k principal components. However, it maximizes the information captured in the first principal components in order to allow the reduction of the dimensionality of the dataset without losing significant amounts of information. The sum of the values of the eigenvalues is equal to 9, since we are considering 9 variables in our model.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.17944459	1.79251034	0.3533	0.3533
2	1.38693425	0.36031007	0.1541	0.5074
3	1.02662419	0.13774769	0.1141	0.6214
4	0.88887649	0.04811525	0.0988	0.7202
5	0.84076124	0.29139603	0.0934	0.8136
6	0.54936522	0.02984569	0.0610	0.8747
7	0.51951953	0.12884350	0.0577	0.9324
8	0.39067602	0.17287756	0.0434	0.9758
9	0.21779846		0.0242	1.0000

For the purposes of our analysis, the most significant eigenvalues are the first three eigenvalues, since we should consider the values higher than 1. We reach **62% of the total variability.**

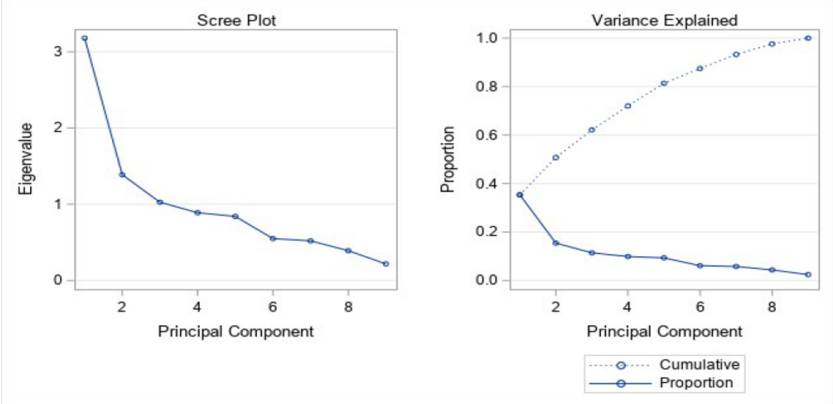
Eigenvectors										
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
a1	imp_Privacy	0.387780	-.475242	-.056173	0.087738	-.222469	-.199938	-.035240	-.345315	0.634796
a2	imp_Security	0.438326	-.307213	-.037279	0.132950	-.200609	-.120286	0.438727	-.105105	-.660442
a3	imp_Feature	0.372392	-.280614	0.329818	-.085259	0.263721	-.221801	-.520681	0.510932	-.125002
a4	imp_Search	0.245297	0.118147	-.220728	0.765066	0.464457	0.243885	0.031613	0.078426	0.097335
a5	imp_Interface	0.282306	0.521162	-.101038	-.166029	0.213258	-.652189	0.320566	0.092918	0.169386
a6	imp_Customization	0.231192	0.206978	0.771938	-.074413	0.220257	0.268116	0.153095	-.395214	0.047468
a7	imp_Integration	0.286830	0.366003	0.154537	0.174845	-.685650	0.231131	0.013173	0.426626	0.153614
a8	imp_MobileInterface	0.362499	0.357328	-.326234	-.142760	-.083034	0.089620	-.581163	-.450731	-.241386
a9	imp_Speed	0.334405	-.104738	-.318664	-.545873	0.232362	0.527437	0.264659	0.229156	0.151462

As explained previously, the eigenvectors represent the direction of the segment, whereas the eigenvalues the data variance magnitude explained by them. Thus, the eigenvector with the highest eigenvalue is the most dominant principal component.

As we can observe in the results, the first principal component presents just positive values, this means that all the variables are positively related one to the others. This is likely due to the presence of the **size effect.**

Size effect is due to the personal value that each respondent assigns to a specific score. In the questionnaire people had to give to the attributes a score between 1 and 7, but the value that each respondent gives to each score is subjective (for example, respondent 1 and respondent 2 can give a different value to score X). It is not possible to reliably measure intangible assets, so it is difficult to understand what 1, 7 and the values in between mean in a personal judgment scale. This happens because opinions cannot be measured objectively: scales will always be

biased by personal judgment. If we were to proceed now with the clustering they would be biased and untruthful, so we have first to proceed with the elimination of the size effect.



3.1 Elimination of Size Effect

In order to eliminate the size effect, we have to create a new dataset and transform the variables related to the scores associated to each respondent into values included between -1 and 1. We calculate the average value of the responses of each respondent, as well as the minimum and maximum score given by each respondent. In this case, the average score of each respondent associated will be equal to 0, the minimum score to -1 and the maximum score to 1.

We need to create an algorithm to transform our data which respects the following rules:

- If OLDi > AVGi then NEWi = $OLDi-AVGi$ / $MAXi-AVGi$
- If OLDi < AVGi then NEWi = $OLDi-AVGi$ / $AVGi-MINi$
- If OLDi = AVGi then NEWi = 0

Where i represents the behavior of one single general respondent.

In this way we can generalize the subjective importance of each attribute for each individual by classifying it: negative attributes will not be important, attributes around 0 will be of average importance, and positive attributes will be considered important.

Now we can proceed with the analysis. First, we have to rerun a principal component analysis to have results unbiased by subjective perceptions (size-effect free).

```
data rest.data_adj; set rest.questionario;
avg_i=mean(of a1-a9);
min_i=min(of a1-a9);
max_i=max(of a1-a9);
array a a1-a9;
array b new1-new9;
do over b;
b=.;
if a>avg_i then b=(a-avg_i)/(max_i-avg_i);
if a<avg_i then b=(a-avg_i)/(avg_i-min_i);
if a=avg_i then b=0;
if a=. then b=0;
end;
```

After calculating the average, the minimum and the maximum of each of the variables, they were transformed into new ones (i.e., “new1-new9”), using the “if” statement in order to set the conditions to restrict our scale and eliminate the biases. After running this code, new names were given to the new variables, by relabeling them:

```
label new1='privacy';
label new2='security';
label new3='feature';
label new4='search';
label new5='interface';
label new6='customization';
label new7='integration';
```

```
label new8='mobileinterface';
label new9='speed';
run;
```

3.2 Final PCA

At the end of this process, the Principal Component Analysis Procedure was run again on the new unbiased variables:

```
proc princomp data=rest.data_adj
out=rest.data_coord_1_adj;
var new1-new9;
run;
```

Removing the biases, we obtain a new correlation matrix, new eigenvalues and new eigenvectors:

Correlation Matrix

		new1	new2	new3	new4	new5	new6	new7	new8	new9
new1	privacy	1.0000	0.4415	0.3125	-.0097	-.2767	-.3320	-.0927	-.1279	0.1242
new2	security	0.4415	1.0000	0.1939	0.0693	-.1264	-.1891	-.0472	-.1466	0.1742
new3	feature	0.3125	0.1939	1.0000	0.0520	-.0702	-.0933	-.1216	0.0257	0.1149
new4	search	-.0097	0.0693	0.0520	1.0000	0.1306	-.0556	-.0133	0.1179	-.0190
new5	interface	-.2767	-.1264	-.0702	0.1306	1.0000	0.0454	0.0360	0.2814	0.0738
new6	customization	-.3320	-.1891	-.0933	-.0556	0.0454	1.0000	0.1153	-.0814	-.1687
new7	integration	-.0927	-.0472	-.1216	-.0133	0.0360	0.1153	1.0000	0.2692	-.1153
new8	mobileinterface	-.1279	-.1466	0.0257	0.1179	0.2814	-.0814	0.2692	1.0000	0.2228
new9	speed	0.1242	0.1742	0.1149	-.0190	0.0738	-.1687	-.1153	0.2228	1.0000

By analyzing the new correlation matrix, we see how the highest positive correlation is still between “security” and “privacy”. And the lowest correlation, this time is negative, is between “customization” and “privacy”.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.04556185	0.54516985	0.2273	0.2273
2	1.50039200	0.43434542	0.1667	0.3940
3	1.06604658	0.04445000	0.1184	0.5124
4	1.02159658	0.12108432	0.1135	0.6260
5	0.90051226	0.08937774	0.1001	0.7260
6	0.81113452	0.09364311	0.0901	0.8161
7	0.71749141	0.23917009	0.0797	0.8959
8	0.47832132	0.01937783	0.0531	0.9490
9	0.45894349		0.0510	1.0000

This new Eigenvalues table differs from the previous one because in this one we have that the first four variables have an eigenvalue greater than 1, and cumulatively explain almost 62,6% of the total variability. Hence, our previous idea is now supported by the data. Looking at the Eigenvectors table it is also evident that there are no more only positive correlations between the variables and the first principal component:

```
proc cluster data=rest.data_coord_1_adj method=ward outtree=rest.data_tree;
var prin1-prin4;
id obs_id;
run;
```

Eigenvectors

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
new1	privacy	0.560561	0.012432	0.225126	0.056110	-.028196	-.063490	0.115280	0.486537	0.613965
new2	security	0.472000	0.064648	0.157797	0.168452	-.038524	0.624040	0.235215	0.043274	-.521203
new3	feature	0.350801	0.153487	-.040537	0.143345	0.758165	-.376050	0.113668	-.300218	-.109136
new4	search	0.003606	0.295780	-.218651	0.808240	-.144377	0.060616	-.406633	-.070474	0.128863
new5	interface	-.289162	0.436966	-.318445	0.063168	0.037125	0.192940	0.714536	-.015329	0.265797
new6	customization	-.351090	-.274770	0.010254	0.094709	0.623526	0.473009	-.185297	0.351496	0.147497
new7	integration	-.232638	0.147560	0.827676	0.123832	0.014132	0.105960	0.055358	-.403454	0.215539
new8	mobileinterface	-.192496	0.626433	0.245268	-.102177	0.075068	-.221612	-.146023	0.551237	-.346408
new9	speed	0.204595	0.451655	-.168079	-.505907	0.077869	0.374370	-.428755	-.275482	0.253020

In this case, after the elimination of the size effect, the relation between attributes in principal component 1 is no longer only positive, but there are some negative correlated attributes. A deeper and more meaningful analysis can be performed now.

If we look in a detailed way at the composition of the first four components, we can analyze their composition by visually displaying the values of attributes, from the less important (negative values) to the most important one. The values in each column (eigenvector) indicate the contribution of each original variable to the particular principal component.

Principal component 1: it is positively influenced by the presence of “security” and “privacy”, whereas it is negatively influenced by the “customization” and the “interface”.

Principal component 2: it is positively influenced by the presence of “mobileinterface”, “speed” and “interface”, whereas it is negatively influenced by the “customization”.

Principal component 3: it is positively influenced by the presence of “integration”, whereas it is negatively influenced by many factors the greatest are “interface” and “search”.

Principal component 4: search – it is positively influenced by the presence of “search”, whereas it is negatively influenced by the “speed” and the “mobile interface”.

4. Clustering

Clustering is a multivariate method used to classify different variables into groups (i.e., clusters) focusing on detecting the similarities between two units in the same group and maximizing the difference between units of different groups.

It is commonly used in exploratory data analysis to discover structures in the data and identify groups of observations that may have similar properties and its results can be visualized using various techniques such as scatter plots, dendrograms, and heatmaps.

Several different methods can be applied to carry out a cluster analysis depending on the data and the desired outcome of the analysis.

4.1 Clustering method and formation

The present analysis utilized Ward's Method to carry out the clustering process using the four principal components previously described with the eigenvalue higher than 1.

Ward's Method is a hierarchical approach that aims to reduce the variance within the clusters while simultaneously maximizing the variance between the clusters. This method is widely used in opinion data analysis. The outcome of applying this method is building different segments from the respondent’s pool, and is particularly useful for identifying compact and well-separated clusters.

On the other hand, the Single-Linkage Method was not chosen for this analysis as it tends to focus on the most densely populated parts of the data and only evaluates the minimum distance between values. This can result in clusters that are elongated and not well-separated, in nested dendrograms, which can hinder the

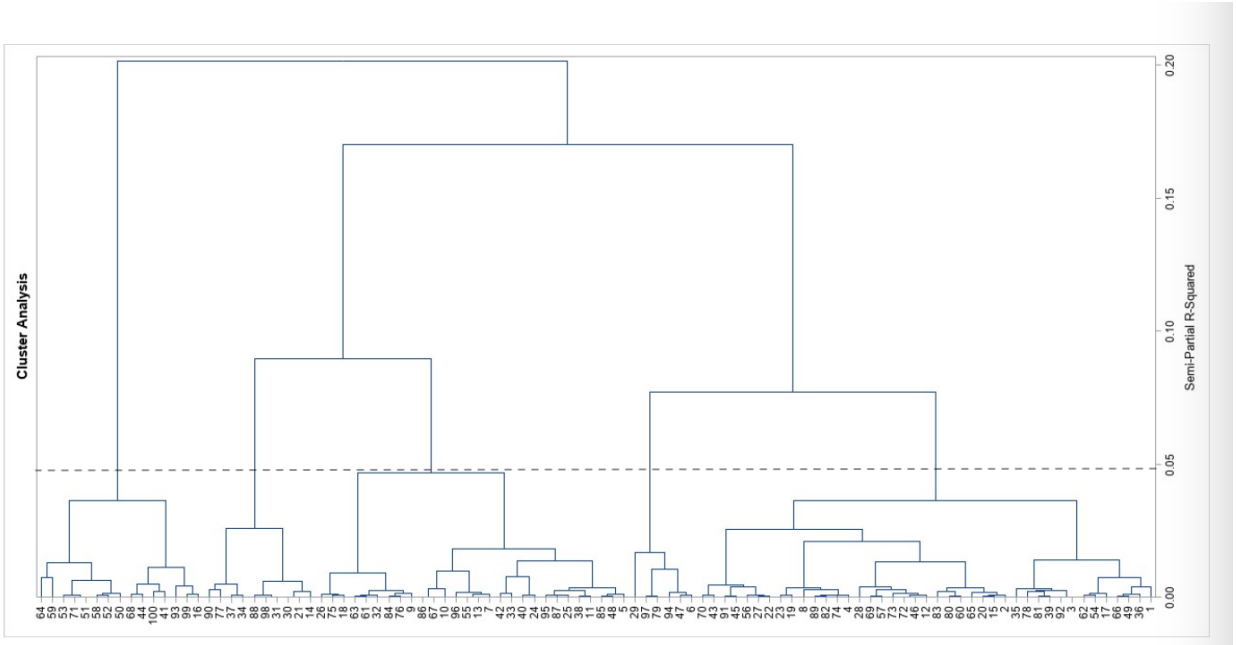
interpretation of the results and are not very useful due to the difficulty to differentiate clusters.

Overall, the use of Ward's Method in conjunction with the selection of appropriate principal components allowed for the identification of well-defined and easily interpretable clusters, which can be used to inform further analysis and decision-making processes. The following codes were used to create the dendrogram and decide the number of clusters:

```
proc cluster data=rest.data_coord_1_adj method=ward outtree=rest.data_tree;
var prin1-prin4;
id obs_id;
run;

proc sgrender data=rest.data_tree template=dendrogram;
run;
```

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.04556185	0.54516985	0.3631	0.3631
2	1.50039200	0.43434542	0.2663	0.6294
3	1.06604658	0.04445000	0.1892	0.8187
4	1.02159658		0.1813	1.0000



Resulting from the previous line of code, is a dendrogram which is used for determining the number of clusters and identifying them, it is useful to consider that the choice of number of clusters is not only a matter of statistical factors but also of final use. The dendrogram is a tree diagram that visually presents the results of the clustering algorithm. On the x-axis we can observe the clusters while the y-axis represents the hierarchical level of aggregation. To identify the number of clusters, we considered not only the statistical factors but also the purpose of the final use of the clusters. Therefore, we cut the dendrogram at the 0,0467 value of the y-axis and obtained five clusters.

To divide the units into five clusters, the proc tree command was run as follows:

```
proc tree data=rest.data_tree ncl=5 out=rest.data_cluster;
id obs_id;
run;
```

4.2 Preparation for the T-test

In order to identify the characteristic attributes of each cluster, a comparison of each variable with the general average is necessary. To accomplish this, the t-test statistical method is employed.

However, the limit of the t-test is the inability to compare the cluster dataset with the original large dataset. To overcome this limitation, an additional cluster is created using the original dataset so that a comparison can be made between the two clusters.

Beforehand, the two datasets had to be sorted by identifier and then merged. In this case, the identifier is used as a key variable. After having created the merged dataset, an additional cluster is created representing the data of the new dataset. Therefore, resulting in 6 total clusters.

It is important to note that the t-test is a parametric test that assumes the normality of the data and equal variances between the groups being compared. If these assumptions are violated, alternative non-parametric tests may be more appropriate.

The procedure explained above was written in code as follows:

```
proc sort data=rest.data_coord_1_adj;
by obs_id;
run;

proc sort data=rest.data_cluster;
by obs_id;
run;

data rest.data_merged;
merge rest.data_coord_1_adj rest.data_cluster;
by obs_id;
run;

data rest.data_fake;
set rest.data_merged;
cluster=6;
run;

data rest.data_appendend;
set rest.data_merged rest.data_fake;
run;
```

4.3 T-test

T-test is a statistical tool used to verify if the average values of two distributions are significantly different from each other. The null hypothesis of the test assumes that the mean values of A and B are equal (H0: $\bar{x}$ A=$\bar{x}$ B). The test is based on the formula ($\bar{x}$ A-$\bar{x}$ B)/σES, where σES is an estimation of the pooled standard deviation, which is calculated as the average between the standard deviations of

the two distributions. An adjusted version has been defined by Satterthwaite, which tried to correct the bias generated by the Pooled formula: in fact, using the average implies using an approximation, as σA is different from σB.

The primary goal of the T-Test in SAS is to analyze clusters and determine the variables that are important in describing them. It is important to look at the obtained p-values for each item of the clusters. Indeed, the p-value indicates the probability for the null hypothesis to be true. For the purpose of this research, a threshold value is established, a so-called significance level, equal to 0,05. If the p-value is equal to or smaller than 0,05, it means that the data distribution is inconsistent with the null hypothesis, therefore the null hypothesis is rejected and the alternative hypothesis is correct, indicating that the variable is a good explanator of the cluster. It follows the SAS codes used to perform the t-test:

```
proc ttest data=rest.data_appendend ;
var new:;
class cluster;
where cluster=1 or cluster=6;
run;

proc ttest data=rest.data_appendend;
var new:;
class cluster;
where cluster=2 or cluster=6;
run;

proc ttest data=rest.data_appendend;
var new:;
class cluster;
where cluster=3 or cluster=6;
run;

proc ttest data=rest.data_appendend;
var new:;
class cluster;
```

```
where cluster=4 or cluster=6;
run;

proc ttest data=rest.data_appendend;
var new:;
class cluster;
where cluster=5 or cluster=6;
run;
```

4.4 Chi-square test

The Chi-Square test is a statistical hypothesis test that is used to state whether the differences between observed outputs and expected outputs are due to an association between the analyzed categorical variables or not.

In this analysis, this test was applied to better understand the composition of the existing clusters and interpret them more detailed through the relation with qualitative variables such as demographic ones: The null hypothesis (H0) assumes independence between the variables, meaning that there is no association between them. The Chi-Square test calculates the difference between the observed and expected outputs and determines the probability of obtaining such results by chance. If the probability is less than the significance level (usually set to 0.05), the null hypothesis is rejected, indicating that there is a significant association between the variables.

The aim was to determine if there was a significant relation between the variables, therefore, to see if the null hypothesis H0 of independence between the variables was verified or not.

The Chi-Square analysis was performed by running the following proc freq code on SAS:

```
proc freq data=rest.data_merged;
table d1*cluster/expected chisq;
run;
```

Frequency
Expected
Percent
Row Pct
Col Pct

Table of d1 by CLUSTER

d1(Gender)	CLUSTER 1	2	3	4	5	Total
Female	4 5.2 4.00 7.69 40.00	22 21.32 22.00 42.31 53.66	17 14.56 17.00 32.69 60.71	7 7.8 7.00 13.46 46.67	2 3.12 2.00 3.85 33.33	52 52.00
Male	6 4.7 6.00 12.77 60.00	19 19.27 19.00 40.43 46.34	10 13.16 10.00 21.28 35.71	8 7.05 8.00 17.02 53.33	4 2.82 4.00 8.51 66.67	47 47.00
Other	0 0.1 0.00 0.00 0.00	0 0.41 0.00 0.00 0.00	1 0.28 1.00 100.00 3.57	0 0.15 0.00 0.00 0.00	0 0.06 0.00 0.00 0.00	1 1.00
Total	10 10.00	41 41.00	28 28.00	15 15.00	6 6.00	100 100.00

Statistics for Table of d1 by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	8	5.5070	0.7023
Likelihood Ratio Chi-Square	8	5.5239	0.7004
Mantel-Haenszel Chi-Square	1	0.0932	0.7601
Phi Coefficient		0.2347	
Contingency Coefficient		0.2285	
Cramer's V		0.1659	

WARNING: 53% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

```
proc freq data=rest.data_merged;
table d2*cluster/expected chisq;
run;
```

Frequency
Expected
Percent
Row Pct
Col Pct

Table of d2 by CLUSTER

d2(Age)	CLUSTER 1	2	3	4	5	Total
18 - 24	3 6 3.00 5.00 30.00	24 24.6 24.00 40.00 58.54	22 16.8 22.00 36.67 78.57	9 9 9.00 15.00 60.00	2 3.6 2.00 3.33 33.33	60 60.00
25 - 34	5 3.3 5.00 15.15 50.00	13 13.53 13.00 39.39 31.71	6 9.24 6.00 18.18 21.43	6 4.95 6.00 18.18 40.00	3 1.98 3.00 9.09 50.00	33 33.00
35 - 44	2 0.4 2.00 50.00 20.00	2 1.64 2.00 50.00 4.88	0 1.12 0.00 0.00 0.00	0 0.6 0.00 0.00 0.00	0 0.24 0.00 0.00 0.00	4 4.00
45 - 60	0 0.3 0.00 0.00 0.00	2 1.23 2.00 66.67 4.88	0 0.84 0.00 0.00 0.00	0 0.45 0.00 0.00 0.00	1 0.18 1.00 33.33 16.67	3 3.00
Total	10 10.00	41 41.00	28 28.00	15 15.00	6 6.00	100 100.00

Statistics for Table of d2 by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	12	20.8627	0.0524
Likelihood Ratio Chi-Square	12	19.5120	0.0769
Mantel-Haenszel Chi-Square	1	0.6444	0.4221
Phi Coefficient		0.4568	
Contingency Coefficient		0.4155	
Cramer's V		0.2637	

WARNING: 70% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

```
proc freq data=rest.data_merged;
table d3*cluster/expected chisq;
run;
```

Frequency Expected Percent Row Pct Col Pct	Table of d3 by CLUSTER					
	CLUSTER					
d3(Region)	1	2	3	4	5	Total
Extra-EU	5 4.1 5.00 12.20 50.00	15 16.81 15.00 36.59 36.59	14 11.48 14.00 34.15 50.00	6 6.15 6.00 14.63 40.00	1 2.46 1.00 2.44 16.67	41 41.00
Intra-EU	0 0.6 0.00 0.00 0.00	2 2.46 2.00 33.33 4.88	4 1.68 4.00 66.67 14.29	0 0.9 0.00 0.00 0.00	0 0.36 0.00 0.00 0.00	6 6.00
Italy	5 5.3 5.00 9.43 50.00	24 21.73 24.00 45.28 58.54	10 14.84 10.00 18.87 35.71	9 7.95 9.00 16.98 60.00	5 3.18 5.00 9.43 83.33	53 53.00
Total	10 10.00	41 41.00	28 28.00	15 15.00	6 6.00	100 100.00

Statistics for Table of d3 by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	8	9.9786	0.2665
Likelihood Ratio Chi-Square	8	11.1890	0.1912
Mantel-Haenszel Chi-Square	1	0.3646	0.5460
Phi Coefficient		0.3159	
Contingency Coefficient		0.3012	
Cramer's V		0.2234	
WARNING: 53% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

```
proc freq data=rest.data_merged;
table d4*cluster/expected chisq;
run;
```

Frequency Expected Percent Row Pct Col Pct	Table of d4 by CLUSTER					
	CLUSTER					
d4(Occupation)	1	2	3	4	5	Total
Employed	2 2.2 2.00 9.09 20.00	9 9.02 9.00 40.91 21.95	4 6.16 4.00 18.18 14.29	5 3.3 5.00 22.73 33.33	2 1.32 2.00 9.09 33.33	22 22.00
Self-employed	0 0.5 0.00 0.00 0.00	3 2.05 3.00 60.00 7.32	2 1.4 2.00 40.00 7.14	0 0.75 0.00 0.00 0.00	0 0.3 0.00 0.00 0.00	5 5.00
Student (employed)	4 2.7 4.00 14.81 40.00	12 11.07 12.00 44.44 29.27	6 7.56 6.00 22.22 21.43	4 4.05 4.00 14.81 26.67	1 1.62 1.00 3.70 16.67	27 27.00
Student (unemployed)	4 4.4 4.00 9.09 40.00	17 18.04 17.00 38.64 41.46	15 12.32 15.00 34.09 53.57	5 6.6 5.00 11.36 33.33	3 2.64 3.00 6.82 50.00	44 44.00
Unemployed	0 0.2 0.00 0.00 0.00	0 0.82 0.00 0.00 0.00	1 0.56 1.00 50.00 3.57	1 0.3 1.00 50.00 6.67	0 0.12 0.00 0.00 0.00	2 2.00
Total	10 10.00	41 41.00	28 28.00	15 15.00	6 6.00	100 100.00

Statistics for Table of d4 by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	16	9.7483	0.8794
Likelihood Ratio Chi-Square	16	11.5776	0.7725
Mantel-Haenszel Chi-Square	1	0.0113	0.9153
Phi Coefficient		0.3122	
Contingency Coefficient		0.2980	
Cramer's V		0.1561	
WARNING: 72% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

```
proc freq data=rest.data_merged;
table d5*cluster/expected chisq;
run;
```

Frequency
Expected
Percent
Row Pct
Col Pct

Table of d5 by CLUSTER

d5(Degree)	CLUSTER 1	2	3	4	5	Total
Bachelor's degree (BA)	3 3.4 3.00 8.82 30.00	12 13.94 12.00 35.29 29.27	11 9.52 11.00 32.35 39.29	5 5.1 5.00 14.71 33.33	3 2.04 3.00 8.82 50.00	34 34.00
High school or college	3 2.6 3.00 11.54 30.00	12 10.66 12.00 46.15 29.27	8 7.28 8.00 30.77 28.57	2 3.9 2.00 7.69 13.33	1 1.56 1.00 3.85 16.67	26 26.00
Master's degree (MA, MBA)	3 3.5 3.00 8.57 30.00	13 14.35 13.00 37.14 31.71	9 9.8 9.00 25.71 32.14	8 5.25 8.00 22.86 53.33	2 2.1 2.00 5.71 33.33	35 35.00
Middle school	0 0.2 0.00 0.00 0.00	2 0.82 2.00 100.00 4.88	0 0.56 0.00 0.00 0.00	0 0.3 0.00 0.00 0.00	0 0.12 0.00 0.00 0.00	2 2.00
PhD and higher	1 0.3 1.00 33.33 10.00	2 1.23 2.00 66.67 4.88	0 0.84 0.00 0.00 0.00	0 0.45 0.00 0.00 0.00	0 0.18 0.00 0.00 0.00	3 3.00
Total	10 10.00	41 41.00	28 28.00	15 15.00	6 6.00	100 100.00

Statistics for Table of d5 by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	16	10.6011	0.8334
Likelihood Ratio Chi-Square	16	12.0327	0.7417
Mantel-Haenszel Chi-Square	1	1.1303	0.2877
Phi Coefficient		0.3256	
Contingency Coefficient		0.3096	
Cramer's V		0.1628	

WARNING: 68% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Based on the p-values obtained, which were all found to be above the conventional threshold of 0.05, the null hypothesis of independence cannot be rejected. Thus, it can be concluded that in the presented case, there is a condition of independence between the examined categorical variables and the obtained clusters. This implies that the examined categorical variables are not statistically associated with the obtained clusters.

5. Cluster description

5.1 General analysis

Five clusters were identified with the analyses carried out in this paper. In this section, the preferences and characteristics of the components of each cluster are analyzed in order to create the most precise sketches possible that can guide the actions of those in charge of deciding which strategies to focus on in order to ensure the best results.

In particular, in outlining the precise characteristics, it was decided to use chi-square values to identify which variables are most significantly related to each consumer group.

From the analysis carried out, a finding emerged that is significant to comment on here, generally referring to all clusters. In fact, it turns out that socio-demographic variables are neither characterizing nor relevant in the distinction into groups of the analyzed consumers.

In particular, it is interesting to note that in all the clusters obtained, the percentage of males and females is approximately equivalent.

Furthermore, the variable concerning the age groups of the questionnaire respondents also shows that consumers belonging to the same age group may have different preferences. It should also be noted that the age factor is to be evaluated in this analysis as potentially biased by the age of those who took part in this work who administered the questionnaire to a majority of their peers. In fact, it is easy to see that respondents are present in greater numbers in the 18-35 age group.

Having analyzed the common characteristics, it is possible to proceed to a more accurate description of the individual clusters.

5.2 Cluster 1

The first cluster identified consists of 10 elements. The individuals are aged between 25 and 34, from Italy and the non-EU area. They are mainly students whose level of education varies from high school to Master's Degree.

Analyzing the preferences expressed in terms of the importance of the various elements taken into consideration, very high expectations emerge, which is why this cluster could be called “Demanding”. In fact, consumers belonging to this cluster show that they attach the highest importance to all but one attribute. In particular, security, feature integration, research reliability, interface quality for both mobile and computer, integration and speed are of the highest importance. The only element that is given an importance of 4 out of 7 is the possibility of customization.

5.3 Cluster 2

The second cluster is considerably larger than the first, comprising 41 elements, again almost equally divided between males and females.

The individuals taken into account belong mainly to the 18 to 24 age group, i.e., they are predominantly students, both employed and unemployed, from Italy and the non-EU area.

In this cluster, two outliers are identified, belonging to the 45-60 age group, which are however not of concern once they are framed in the previous discourse that does not consider demographic variables significant.

Compared to the first clusters, the importance attributed varies significantly. In particular, the highest importance is attributed to privacy, security, search and speed.

On the contrary, other variables, such as features integration, interface, customization, and integration with other devices revealed an attribution of medium importance, hence not priority.

5.4 Cluster 3

The third identified cluster is also quite numerous, counting a total of 28 elements, of which 17 females, 10 males, and 1 identified as 'other'.

This cluster is also primarily composed of individuals aged between 18 and 24, thus predominantly unemployed students from outside the European Union, with a level of education predominantly equivalent to a Bachelor’s Degree.

In this cluster, there is a slight decrease in the importance attributed to privacy, which drops to a score of 6. Security, search accuracy, speed, and both mobile and computer interface remain of the utmost importance.

Integration of features, customization and integration with other devices, on the other hand, have low average importance values, in a range between 3 and 5.

5.5 Cluster 4

The fourth cluster consists of 15 individuals, equally distributed between males and females, belonging to the age group of 18 to 24 years. Again, they are students or unemployed, coming from Italy and the non-EU area, mainly with an education level equivalent to a Master's Degree.

In this cluster there is an additional element that would be considered an outlier if it were not primarily identified on demographic grounds.

The characteristics highlighted by the analysis of this consumer group are very interesting, departing from a trend that seemed to be common to the other clusters, that makes it possible to call this cluster “Unsafe”. The importance of privacy is in fact given a very low score, equivalent to 2, while security is given a score of 4. Also surprising is the score, varying between 3 and 5, given to the speed of the browsers, which appears to be in the other clusters part of the priorities. Finally, being rated as less important than the other clusters is feature integration, with the least importance.

On the other hand, the variables concerning the interface, device integration, and search reliability are prioritized.

5.6 Cluster 5

To conclude, the last cluster identified is extremely small in size. The 6 individuals composing it, 4 males and 2 females, are aged between 18 and 34, and are students or unemployed Italians with an education level varying between Bachelor's Degree and Master's Degree.

In this consumer group, priorities such as privacy, security and speed are given top priority. In contrast to the previous clusters, however, there is an increase in the importance attributed to customization, which rises to a score of 6, and a decrease in the importance of search reliability, which falls to 4.

Finally, interface and integration are given a rather high, if not maximum, importance.

Conclusion

In conclusion, the tools provided by SAS made it possible to analyze the distribution of consumer preferences in terms of the importance of the main attributes of a browser.

In particular, one must take into consideration the database on which the research was developed. In fact, as was previously noted, the respondents are predominantly rather young students, aged between 18 and 34.

For this reason, one must consider that the use made of browsers is influenced by the lifestyle of the respondents and their occupations. I.e., a student is more likely to use browsers for school research; a young person is more likely to use a mobile phone than an older person.

The purpose of evaluating the results obtained is primarily to understand which characteristics it is most important to enhance, which to maintain, and which can be given a lower priority. In this way, it is possible to obtain guidelines for managers who have to orient their companies towards the acquisition of new customers and the retention of those already loyal. In particular, this analysis makes it possible to define the quality levels required by consumers in the different clusters identified.

Although five clusters were identified, it should be noted that priority is given to the demands of the most numerous ones, i.e., clusters number 2, 3, and 4.

The analysis shows that the highest priority should be given to elements such as privacy, security, speed and search reliability.