SMARTPHONE CUSTOMERS RESEARCH

Group members: Topi Sidorela, jose haro, gioele fabozzi, doménica cabezas

Table of Contents

1. Introduction

1.1 Aim of the Analysis

According to Statista, by 2021 there were 6,259 million of smartphone users[1] in the world and is forecast to further grow by several hundred million in the next few years. Therefore, the aim of this research is to analyze the users' preferences about smartphones' brands and features as the smartphone market has still a lot of growth potential.

1.2 Survey Description

The survey has been carried out through Google Forms, and it gathered 23 questions in total with different type of questions like multiple choice, likert scale, rating, and drop-down questions.

By the end of the survey, we collected a total of 147 answers from subjects who owned a smartphone. The survey was distributed only in English, and it was composed by 5 socio-demographic questions, 6 behavioral questions, and 12 quantitative importance questions measured with a scale going from 1 to 10. The survey questions divided by section are the following:

Socio-demographic Questions

1. Gender*
Male
Female
Prefer not to say
Other

2. Age group*
15-19
20-24
25-29
30-34
35-40
over 40

3. Country *
[]

4. Education Level*
School
Undergraduate
Postgraduate
PhD

5. Working status*

[1] https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

Full-time worker
Part-time worker
Student
Retired
Others

Behavioral Questions

6. Average annual salary (euros)*
I don't have a salary yet.
0-15,000
15,000-28,000
28,000-35,000
35,000 - 50,000
More than 50,000

7. Amount spent on a mobile phone (euros)*
200-400
400-600
600-800
800-1,000
More than 1,000

8. Which brand of Smartphone did you purchase last time? *
Apple
Samsung
Xiaomi
Huawei
One plus
OPPO
Sony
Vivo
Motorola
Lenovo
Other

9. What is your satisfaction with your current smartphone? *
Very satisfied
Satisfied
Neutral
Dissatisfied
Very dissatisfied

10. What is the primary usage of your phone?*
Business

Social
Gaming
Photography

11. Will you upgrade you smartphone to the latest model of the same brand?*
Yes
No
I don´t know

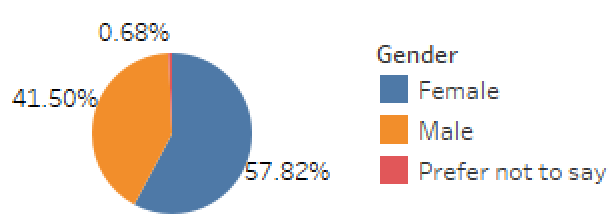
The CORE COMPONENT: IMPORTANT QUESTIONS
This question module included questions about the significance of 12 quantitative variables. Respondents were asked to rate the significance of the following questions on a scale of 1 to 10, being 1 as Not Important and 10 as Really Important:

12. How important is it to you that a smartphone has wireless charging?
13. How important is it to you that a smartphone has finger sensor?
14. How important is it to you that the cell phone has multi-camera?
15. How important is it to you that a smartphone has OLED display?
16. How important is it to you that a smartphone has a simple interface?
17. How important is it to you that a smartphone comes in different colors?
18. How important is it to you that a smartphone is able to fit in your pocket?
19. How important is it to you that a smartphone has a good gaming capacity?
20. How important is it to you that a smartphone has an elegant design?
21. How important is it to you the brand when you purchase a smartphone?
22. How important is it to you that the smartphone has a facial unlocking system?
23. How important is it to you that the smartphone to be compatible with other devices (PC, smartwatch, others)?

2. Preliminary Analysis
The collected answers were then analyzed using the SAS software.

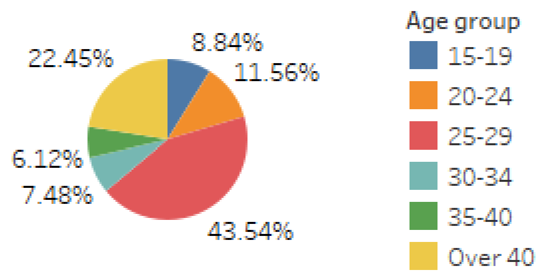
2.1 Sample Description

1. Gender



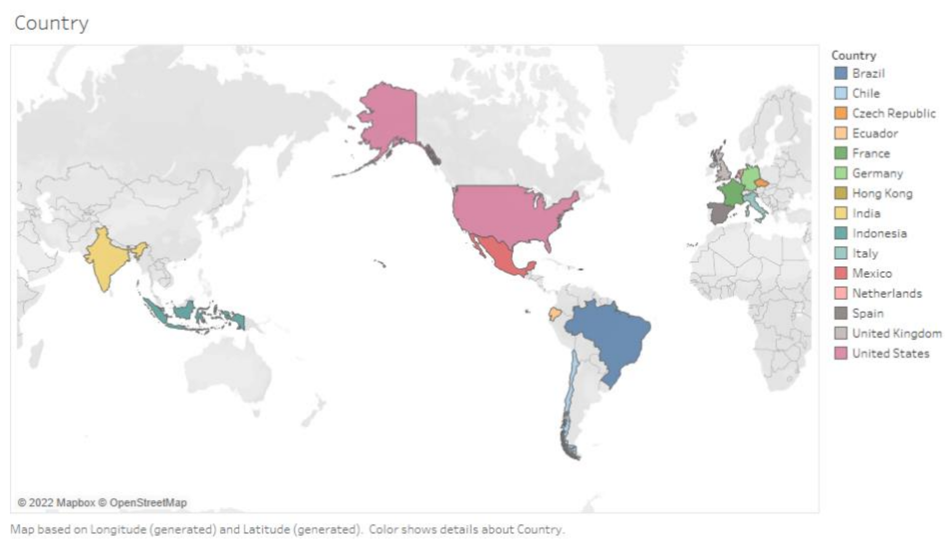
As seen in the previous graph, the 58% of the participants are female, and 42% are male. Additionally, there was only 1 person who preferred not to mention its gender.

2. Age group

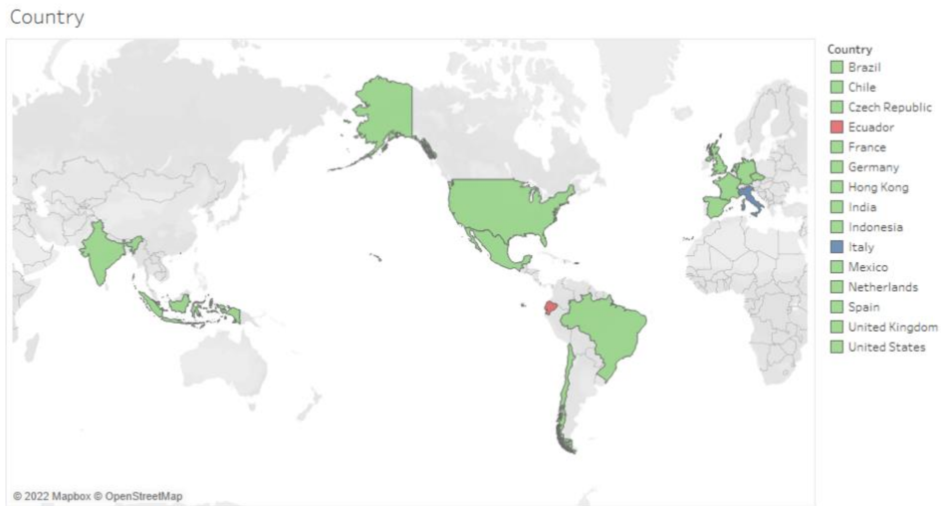


The group with most people belongs to the 25-29 year old with 43.54% of the participants, followed by the group of over 40 years old and the groups of 20-24 years with 22.45% and 11.56%, respectively.

3. Country



We got answers from 15 different countries found in 3 different continents; however, when looking at the distribution, we found that 47.3% of the participants were located in Italy, 33.1% in Ecuador, and 19.6% in other countries. Therefore, rather than distributing the countries by continent, we divided into Italy, Ecuador and Others, as seen in the graph below:



Map based on Longitude (generated) and Latitude (generated). Color shows details about Country.

The distribution of the participants between the 2 principal countries is due to the fact that the authors of this document belong to Italy and Ecuador.

4. Education Level

Education level	
Phd	1
Postgraduate	86
School	21
Undergraduate	39

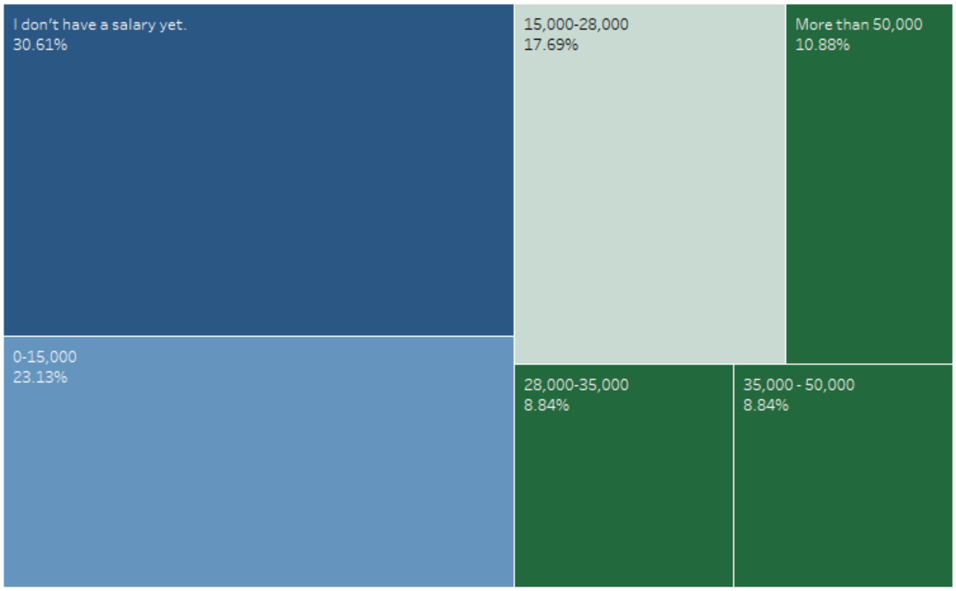
The majority of the participants have Postgraduate studies, while only 1 participant has a PhD. Then, 39 participants have a Bachelor Degree and 21 of them graduated from highschool.

5. Working status

Working status	
Full-time worker	89
Other	8
Part-time worker	9
Retired	2
Student	39

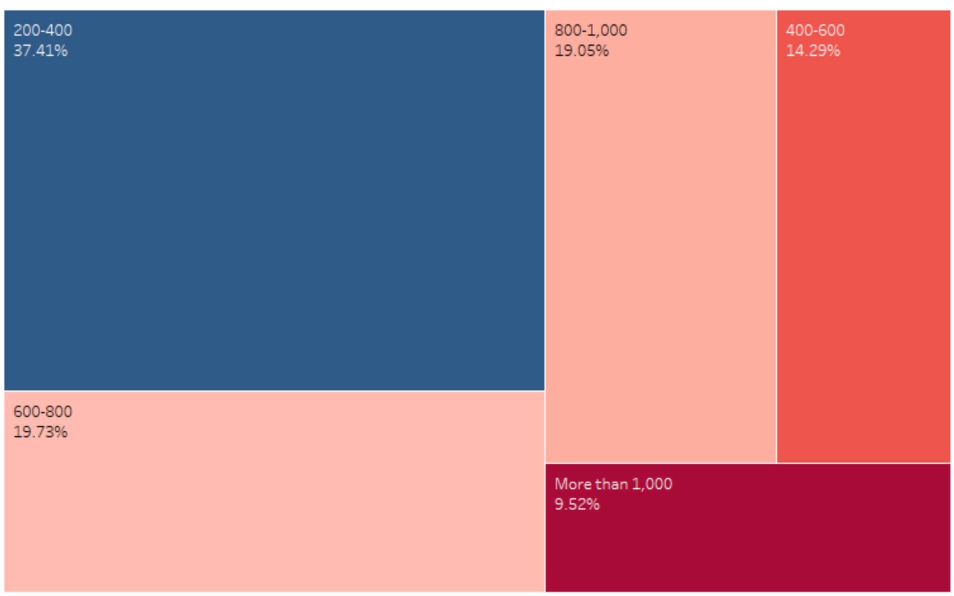
In relation to the previous question, we can see that 89 participants are full-time workers, followed by 39 participants are currently studying. We can also see that 8 people have other working status that could be self-employment, free-lancer, or unemployment.

6. Average annual salary (euros)



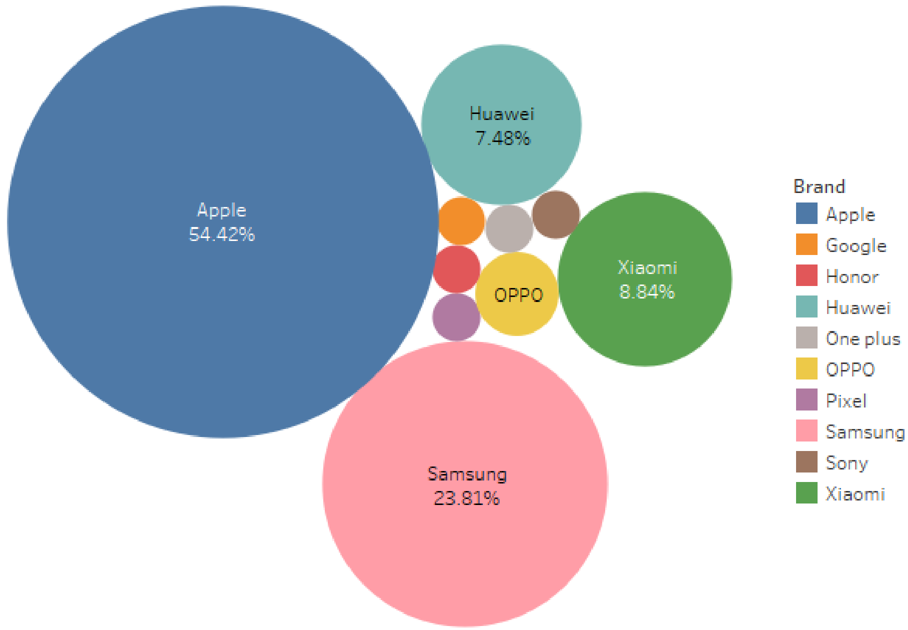
As the second largest group according to working status belong to students, we can see that 31% of our participants don´t have a salary yet. Additionally, we can see that 23% of the people earn from 0 to 15,000 euros per year, and approximately 19% earn more than 35,000 euros each year.

7. Amount spent on a mobile phone (euros)



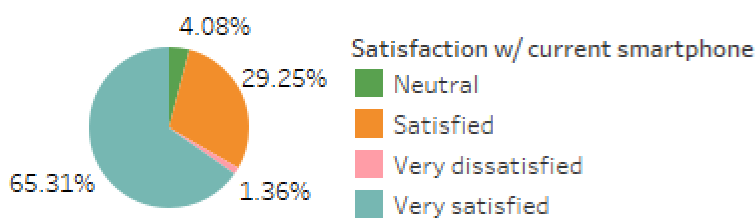
There are smartphones in every price range; however, 37% of the participants agreed that they spend between 200 and 400 euros on a mobile phone. Around 39% of the participants spend between 600 and 1,000 euros, and only 10% spend more than 1,000 euros.

8. Which brand of Smartphone did you purchase last time?



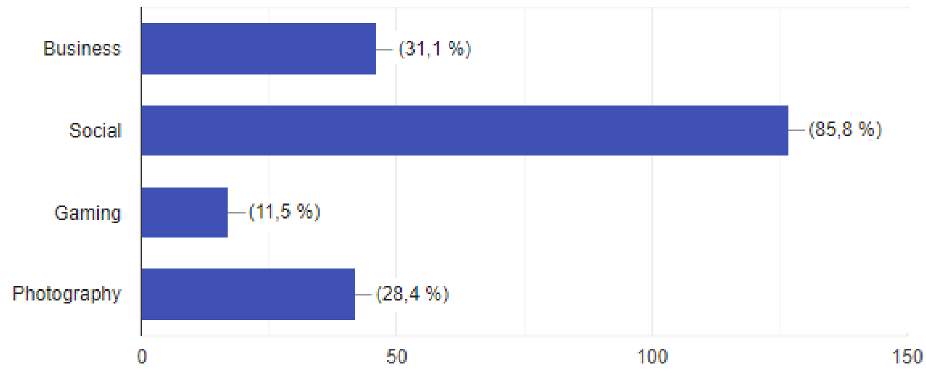
As the worldwide market suggests, the majority of the people have an Apple smartphone, followed by Samsung with the 24% of the participants. Nevertheless, truth the last years we have seen an increase in the market from the Asian brands such as Huawei and Xiaomi.

9. What is your satisfaction with your current smartphone?



As seen in the previous graph, the grand majority feels very satisfied with they current smartphone, 29% are satisfied, and around 5% feel neutral or dissatisfied with their smartphone.

10. What is the primary usage of your phone?



It is important to mention that in this specific question, participants were allowed to choose more than one answer. Therefore, we can see that most of the people use their device for social media and social interaction, followed by business and photography. Only an small percentage (11.5%) of the participants use their device for gaming.

11. Important questions

Question	Min	1Q	Median	Mean	3Q	Max
How important is it to you that a smartphone has wireless charging?	1.00	3.00	7.00	5.97	8.00	10.00
How important is it to you that a smartphone has finger sensor?	1.00	6.00	8.00	7.28	10.00	10.00
How important is it to you that the smartphone has multi-cameras?	1.00	6.00	8.00	7.52	10.00	10.00
How important is it to you that a smartphone has an OLED display?	1.00	5.50	8.00	7.18	9.00	10.00
How important is it to you that a smartphone has a simple interface?	1.00	7.50	9.00	8.29	10.00	10.00
How important is it to you that a smartphone comes in different colors?	1.00	3.00	6.00	5.46	7.00	10.00
How important is it to you that a smartphone is able to fit in your pocket?	1.00	7.00	9.00	8.29	10.00	10.00
How important is it to you that a smartphone has a good gaming capacity?	1.00	2.50	6.00	5.31	8.00	10.00
How important is it to you that a smartphone has an elegant design?	1.00	7.00	8.00	7.92	10.00	10.00
How important is it to you the brand when you purchase a smartphone?	1.00	7.00	8.00	7.71	10.00	10.00
How important is it to you that the smartphone has a facial unlocking system?	1.00	6.00	8.00	7.48	10.00	10.00
How important is it to you that the smartphone to be compatible with other devices (PC, smartwatch, others)?	1.00	8.00	10.00	8.81	10.00	10.00

From the 12 core questions, we got a summary statistics from the answers gotten from each questions. Obviously, we can see that the minimum and the maximum belong to the same minimum and the maximum from the scale.

Apparently, the participants of the survey on average find the color of the device as a feature for which they are indiferent, meaning they don't find it important or not important. On the other hand, on average, people think that the ability of an smartphone to be compatible with other devices is one of the most important features.

Finally, we can observe that all of the features are more or less important to the participants; however, we can see that on average, participants agree that the wireless charging is one of the least important features of an smartphone in comparison with the other qualitative variables.

3. Principal Component Analysis (PCA)

3.1 PCA Analysis

To have a first look at the database's contents, we utilize the CONTENTS procedure,

```
proc contents position data=project.Survey; run;
```

This procedure creates summary information about a dataset's contents, such as:

- The names, types, and attributes of the variables
- The number of observations in the dataset
- The number of variables in the dataset
- Creation date of the dataset

In this step we rename the so called important questions and print the database's contents again.

```
data project.New_Survey
(rename=(How_important_is_it_to_you_that_ = Q_1
            How_important_is_it_to_you_that0 = Q_2
            How_important_is_it_to_you_that1 = Q_3
            How_important_is_it_to_you_that2 = Q_4
            How_important_is_it_to_you_that3 = Q_5
            How_important_is_it_to_you_that4 = Q_6
            How_important_is_it_to_you_that5 = Q_7
            How_important_is_it_to_you_that6 = Q_8
            How_important_is_it_to_you_that7 = Q_9
            How_important_is_it_to_you_the_b = Q_10
            How_important_is_it_to_you_that8 = Q_11
            How_important_is_it_to_you_that9 = Q_12));
set project.Survey;
RUN;
```

```
proc contents position data=project.New_survey; run;
```

The scale used in our questionnaire is from 1 to 10. However, when customers are given opinion surveys in which they are asked to provide a rating on a scale to measure their satisfaction with certain products or services, their perceptions about the measurement scale highly impacts their judgment of intangible notions.

Furthermore, the use of this scale results in a correlation matrix that is almost always formed of coefficients with a positive sign, since respondents tend to offer feedback based on their personal and latent average rating. This is referred to as the "Size Effect."

The PRINCOMP procedure in SAS is used to examine the Size effect by correlation structure and perform a principal component analysis (PCA).

The primary goal of PCA is to minimize the complexity of the interrelationships between a potentially large number of observed variables to a relatively small number of linear combinations of them, known as principle components.

Principal components are weighted linear combinations of the variables where the weights are chosen to account for the largest amount of variation in the data. The total number of principal components is the same as the number of input variables.

We performed the PCA in SAS by using the following code:

```
proc princomp data=project.New_survey;
var Q_1-Q_12;
run;
```

The tables below display the PROC PRINCOMP output, beginning with simple statistics and followed by the correlation matrix.

More in detail, the first section reports the number of observations and variables used along with the simple summary stats (mean and standard deviation) for each variable.

The SAS System

The PRINCOMP Procedure

Observations	147
Variables	12

Simple Statistics

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_10	Q_11	Q_12
Mean	5.965986395	7.278911565	7.517006803	7.176870748	8.292517007	5.455782313	8.285714286	5.306122449	7.918367347	7.714285714	7.482993197	8.809523810
StD	3.119569179	2.766686146	2.612807605	2.536583508	1.821526096	2.772742306	1.951465905	2.992255403	2.249840798	2.443890902	2.760667334	1.917237358

The second section of the printout gives the Pearson correlation matrix for the twelve variables.

Correlation Matrix

		Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_10	Q_11	Q_12
Q_1	How important is it to you that a smartphone has wireless chargi	1.0000	0.3757	0.3879	0.3219	0.3742	0.2805	-.1109	0.2066	0.1313	0.3787	0.3200	0.0699
Q_2	How important is it to you that a smartphone has finger sensor?	0.3757	1.0000	0.2964	0.3736	0.2813	0.2503	-.0364	0.2056	0.1159	0.1689	0.1715	0.0475
Q_3	How important is it to you that the smartphone has multi-cameras	0.3879	0.2964	1.0000	0.4388	0.2875	0.2603	0.0004	0.3668	0.3265	0.2957	0.4171	0.1784
Q_4	How important is it to you that a smartphone has an OLED display	0.3219	0.3736	0.4388	1.0000	0.3712	0.2514	0.1613	0.3321	0.2522	0.3397	0.4914	0.2337
Q_5	How important is it to you that a smartphone has a simple interf	0.3742	0.2813	0.2875	0.3712	1.0000	0.2040	0.1748	0.0878	0.3351	0.4159	0.4457	0.2083
Q_6	How important is it to you that a smartphone comes in different	0.2805	0.2503	0.2603	0.2514	0.2040	1.0000	0.0568	0.2382	0.4946	0.3034	0.3451	0.0834
Q_7	How important is it to you that a smartphone is able to fit in y	-.1109	-.0364	0.0004	0.1613	0.1748	0.0568	1.0000	0.1433	0.2456	0.1896	0.1166	0.2215
Q_8	How important is it to you that a smartphone has a good gaming c	0.2066	0.2056	0.3668	0.3321	0.0878	0.2382	0.1433	1.0000	0.2072	0.2003	0.2921	0.1487
Q_9	How important is it to you that a smartphone has an elegant desi	0.1313	0.1159	0.3265	0.2522	0.3351	0.4946	0.2456	0.2072	1.0000	0.4517	0.4287	0.1980
Q_10	How important is it to you the brand when you purchase a smartph	0.3787	0.1689	0.2957	0.3397	0.4159	0.3034	0.1896	0.2003	0.4517	1.0000	0.5952	0.3815
Q_11	How important is it to you that the smartphone has a facial unlo	0.3200	0.1715	0.4171	0.4914	0.4457	0.3451	0.1166	0.2921	0.4287	0.5952	1.0000	0.4200
Q_12	How important is it to you that the smartphone to be compatible	0.0699	0.0475	0.1784	0.2337	0.2083	0.0834	0.2215	0.1487	0.1980	0.3815	0.4200	1.0000

The last section of the output provides the eigenvalues and eigenvectors for each axis. The eigenvectors and eigenvalues of a covariance (or correlation) matrix represent the “core” of a PCA.

The eigenvectors indicate the relative importance of each variable within the individual axes, hence, they determine the directions of the new feature space. In addition, the new variables (PCs) have a variance equal to their correspoding eigenvalue.

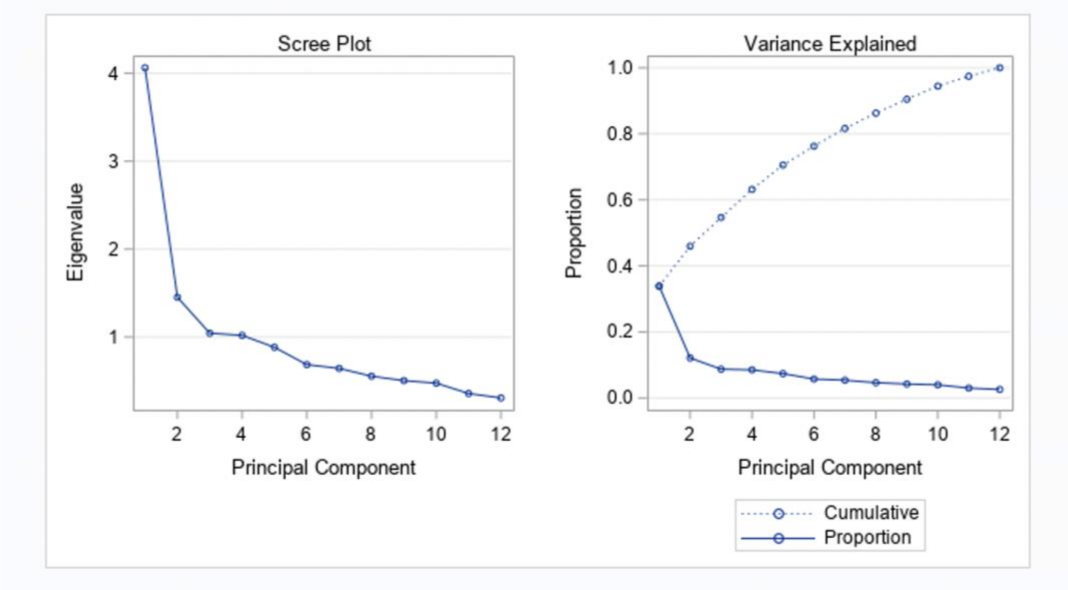
We can state that, the first principal component accounts for about 33,86% of the total variance, the second principal component accounts for about 12,11%, and the third principal component accounts for about 8,70%. The eigenvalues indicate that some components provide a good summary of the data: the first four PCs explain 63% of the total variance in the data. Subsequent components account for less than 7% each.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.06269177	2.60953101	0.3386	0.3386
2	1.45316076	0.40856947	0.1211	0.4597
3	1.04459129	0.02442411	0.0870	0.5467
4	1.02016718	0.13604095	0.0850	0.6317
5	0.88412623	0.19733330	0.0737	0.7054
6	0.68679293	0.04263245	0.0572	0.7626
7	0.64416049	0.08861563	0.0537	0.8163
8	0.55554486	0.05023353	0.0463	0.8626
9	0.50531133	0.02873084	0.0421	0.9047
10	0.47658049	0.11784211	0.0397	0.9444
11	0.35873838	0.05060411	0.0299	0.9743
12	0.30813428		0.0257	1.0000

Moreover, we can see that the first principal component only has positive values: all variables are positively correlated, so if one grows, so does the other. This result is caused by the presence of the size effect. Therefore, this analysis cannot be considered reliable enough to base decisions and conclusions.

Eigenvectors													
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
Q_1	How important is it to you that a smartphone has wireless chargi	0.280397	-.413938	-.215935	-.161773	-.017071	-.108489	0.560956	0.109509	0.354878	0.182418	0.271766	0.324711
Q_2	How important is it to you that a smartphone has finger sensor?	0.227635	-.441388	-.035339	0.102037	0.429832	0.558775	-.097765	0.190914	-.197472	-.345128	-.158573	0.124310
Q_3	How important is it to you that the smartphone has multi-cameras	0.318347	-.225954	0.084247	0.216644	-.218536	-.379848	-.328732	0.572782	0.235675	-.003259	-.269556	-.192790
Q_4	How important is it to you that a smartphone has an OLED display	0.337330	-.107984	-.090399	0.324292	0.166287	-.037919	-.401152	-.533702	0.330670	0.005747	0.392464	-.151519
Q_5	How important is it to you that a smartphone has a simple interf	0.309879	0.016525	-.340237	-.237741	0.393801	-.302329	-.056356	0.021183	-.518051	0.403788	-.023760	-.222097
Q_6	How important is it to you that a smartphone comes in different	0.273857	-.057216	0.544517	-.362236	-.030828	0.351083	0.025845	-.155788	0.191566	0.431131	-.233178	-.260623
Q_7	How important is it to you that a smartphone is able to fit in y	0.115649	0.526622	0.129792	0.275226	0.596901	-.089891	0.247390	0.133954	0.334592	0.023666	-.213278	0.123219
Q_8	How important is it to you that a smartphone has a good gaming c	0.235612	-.079940	0.387907	0.580552	-.183947	-.086761	0.408353	-.118692	-.465430	0.048871	0.082705	-.050255
Q_9	How important is it to you that a smartphone has an elegant desi	0.305864	0.256869	0.411851	-.332724	0.022579	-.097450	-.212645	0.233232	-.164215	-.241210	0.519559	0.308912
Q_10	How important is it to you the brand when you purchase a smartph	0.354190	0.203524	-.189423	-.226924	-.147590	0.001251	0.330700	-.098851	0.056964	-.585367	-.072198	-.503709
Q_11	How important is it to you that the smartphone has a facial unlo	0.384052	0.147166	-.151074	-.044607	-.275920	-.059140	-.136672	-.345910	-.079124	-.043332	-.496078	0.577788
Q_12	How important is it to you that the smartphone to be compatible	0.219120	0.393428	-.359097	0.215876	-.308161	0.535736	-.041544	0.313163	-.007528	0.304935	0.214005	-.032173

Finally, the PROC PRINCOMP procedure generates a scree plot, which shows the proportion of variance explained by each component.



The panel displays two graphs that exhibit the numbers from the table "Eigenvalues of the Correlation Matrix." The scree plot, which is a graph of the eigenvalues, is shown on the left. The total of the eigenvalues equals the number of variables in the analysis, which is 12.

If you divide each eigenvalue by 12, you obtain the proportion of variance that each principal component explains. The graph on the right depicts the proportions and the cumulative proportions.

3.2 Removal of Size Effect

When doing this type of study, the style of the questionnaire, particularly the adjectives used, should be carefully examined. In addition, it is recommended to use a scale ranging from 1 to 10. Given the wide range of separation between adequacy and insufficiency (6-5) votes, the evident advantage of this type of scale is the simplicity with which a vote may be communicated.

As a result of the presence of the "size effect," there is a cluster of subjects who desire all qualities and one who does not want anyone in particular. This outcome is unacceptable since a market segmentation should create groups which differ among them for the mix of favored features.

This happens because respondents tend to provide their ratings by referring to a personal and latent average vote, the correlation matrix of the attributes is always composed of coefficients of a positive sign.

One potential solution is to transform the original data and therefore eliminate the size effect. Given a matrix of n rows and p columns, where the rows are the respondents and the columns are the (quantitative) rate on features of an opinion survey, three important values for recoding will be generated for each respondent: the highest rate (Max), the lowest rate (Min), and the average rate (Avg).

According to the recoding method:

$$
\begin{aligned}
k_{ij} &= 0 && \text{if} \quad x_{ij} = x_{avg} \\
k_{ij} &= \frac{(x_{ij} - x_{min})}{(x_{avg} - x_{min})} && \text{if} \quad x_{ij} < x_{avg} \\
k_{ij} &= \frac{(x_{ij} - x_{max})}{(x_{max} - x_{avg})} && \text{if} \quad x_{ij} > x_{avg}
\end{aligned}
$$

The solution of size effect is represented by a new scale having values having values [-1,+1] as shown in the graph below.

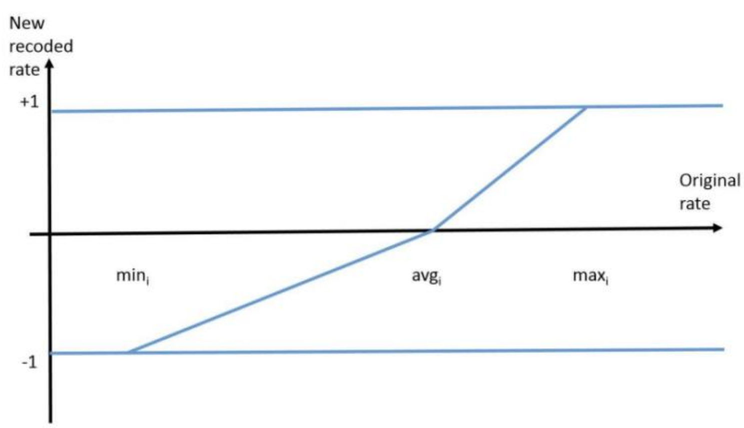


Fig. 1 Example of a recoding function

An important property of this type of scaling transformation is that all respondents, in the new system, have a vector of opinions between -1 and +1, where 0 is the correspondent value of the average of each unit, +1 is max and -1 is min.

Below, we have reported the SAS code used to solve the size effect problem:

```
data data project.New_survey_new; set project.New_survey;
avgi=mean (of Q_1-Q_12);
mini=min (of Q_1-Q_12);
maxi=max (of Q_1-Q_12);
array p1 Q_1-Q_12;
array p2 new_1-new_12;
do over p2;
if p1>avgi then p2=(p1-avgi)/(maxi-avgi);
if p1<avgi then p2=(p1-avgi)/(avgi-mini);
if p1=avgi then p2=0;
if p1=. then p2=0;
end;
label new_1='wireless charging';
label new_2='finger sensor';
label new_3='multi cameras';
```

```
label new_4='OLED display';
label new_5='simple interface';
label new_6='different colors';
label new_7='fits in pocket';
label new_8='gaming';
label new_9='elegant design';
label new_10='brand';
label new_11='face ID';
label new_12='compatibility';
run;
```

The CORR procedure produces Pearson correlation coefficients of the new numeric variables.

```
proc corr data=project.New_survey_new;
var new_1-new_12;
run;
```

The SAS System

The CORR Procedure

12 Variables:	new_1 new_2 new_3 new_4 new_5 new_6 new_7 new_8 new_9 new_10 new_11 new_12

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
new_1	147	-0.24063	0.71566	-35.37328	-1.00000	1.00000	wireless charging
new_2	147	0.16504	0.73668	24.26139	-1.00000	1.00000	finger sensor
new_3	147	0.25564	0.67844	37.57963	-1.00000	1.00000	multi cameras
new_4	147	0.10565	0.65117	15.53039	-1.00000	1.00000	OLED display
new_5	147	0.47573	0.52389	69.93296	-1.00000	1.00000	simple interface
new_6	147	-0.39058	0.66523	-57.41454	-1.00000	1.00000	different colors
new_7	147	0.42064	0.65376	61.83462	-1.00000	1.00000	fits in pocket
new_8	147	-0.41447	0.68467	-60.92654	-1.00000	1.00000	gaming
new_9	147	0.33416	0.63179	49.12201	-1.00000	1.00000	elegant design
new_10	147	0.29427	0.65713	43.25727	-1.00000	1.00000	brand
new_11	147	0.24704	0.70363	36.31519	-1.00000	1.00000	face ID
new_12	147	0.64678	0.57007	95.07611	-1.00000	1.00000	compatibility

The following statement, PROC MEANS produces four basic statistics (N, Min, Max, Mean) for each numeric variables in the last created dataset.

```
proc means data=project.New_survey_new mean min max n;
var new: ;
run;
```

The SAS System

The MEANS Procedure

Variable	Label	Mean	Minimum	Maximum	N
new_1	wireless charging	-0.2406346	-1.0000000	1.0000000	147
new_2	finger sensor	0.1650435	-1.0000000	1.0000000	147
new_3	multi cameras	0.2556437	-1.0000000	1.0000000	147
new_4	OLED display	0.1056489	-1.0000000	1.0000000	147
new_5	simple interface	0.4757344	-1.0000000	1.0000000	147
new_6	different colors	-0.3905751	-1.0000000	1.0000000	147
new_7	fits in pocket	0.4206437	-1.0000000	1.0000000	147
new_8	gaming	-0.4144662	-1.0000000	1.0000000	147
new_9	elegant design	0.3341634	-1.0000000	1.0000000	147
new_10	brand	0.2942671	-1.0000000	1.0000000	147
new_11	face ID	0.2470421	-1.0000000	1.0000000	147
new_12	compatibility	0.6467762	-1.0000000	1.0000000	147

After the removing of the size effect present in our survey, we use again the PRINCOMP procedure.

PCA is an orthogonal linear transformation that transfers data to a new coordinate system such that the data's greatest variance by any projection falls on the first coordinate (first principal component), the second greatest variance falls on the second coordinate (second principal component), and so on.

```
proc princomp data=project.New_survey_new out=project.coord_new;
var new_1-new_12;
run;
```

The SAS System

The PRINCOMP Procedure

Observations	147
Variables	12

Simple Statistics

	new_1	new_2	new_3	new_4	new_5	new_6	new_7	new_8	new_9	new_10	new_11	new_12
Mean	-.2406345748	0.1650434738	0.2556437373	0.1056489094	0.4757344090	-.3905750810	0.4206436765	-.4144662349	0.3341633658	0.2942671451	0.2470421395	0.6467762420
StD	0.7156588297	0.7366806080	0.6784448132	0.6511724799	0.5238855020	0.6652253674	0.6537568601	0.6846650737	0.6317880145	0.6571313428	0.7036289389	0.5700654773

Moreover, after the elimination of the size effect, we can now observe in the correlation matrix, coefficients with a negative sign for e.g., in new_1 we can notice that different colors, fits in pocket, elegant design, face ID and compatibility have a negative sign.

Correlation Matrix

		new_1	new_2	new_3	new_4	new_5	new_6	new_7	new_8	new_9	new_10	new_11	new_12
new_1	wireless charging	1.0000	0.1148	0.0744	0.0172	0.0505	-.0301	-.3009	-.0655	-.1170	0.0343	-.0342	-.1554
new_2	finger sensor	0.1148	1.0000	0.0026	0.1407	0.0427	-.1462	-.0925	-.0970	-.1698	-.1401	-.1491	-.0219
new_3	multi cameras	0.0744	0.0026	1.0000	0.1978	-.0205	-.0262	-.1241	0.0928	-.0007	0.0120	0.2030	-.0340
new_4	OLED display	0.0172	0.1407	0.1978	1.0000	0.1153	-.2425	-.0391	0.0372	-.0260	0.0365	0.2219	0.0910
new_5	simple interface	0.0505	0.0427	-.0205	0.1153	1.0000	-.1222	0.0316	-.1969	-.0328	0.1407	0.1310	0.0320
new_6	different colors	-.0301	-.1462	-.0262	-.2425	-.1222	1.0000	-.0728	0.0019	0.2295	-.0294	0.0494	-.1729
new_7	fits in pocket	-.3009	-.0925	-.1241	-.0391	0.0316	-.0728	1.0000	-.0151	0.0061	-.0301	-.0048	0.1092
new_8	gaming	-.0655	-.0970	0.0928	0.0372	-.1969	0.0019	-.0151	1.0000	0.0397	-.1091	0.0515	-.1276
new_9	elegant design	-.1170	-.1698	-.0007	-.0260	-.0328	0.2295	0.0061	0.0397	1.0000	0.1628	0.1196	-.0204
new_10	brand	0.0343	-.1401	0.0120	0.0365	0.1407	-.0294	-.0301	-.1091	0.1628	1.0000	0.3700	0.1133
new_11	face ID	-.0342	-.1491	0.2030	0.2219	0.1310	0.0494	-.0048	0.0515	0.1196	0.3700	1.0000	0.1908
new_12	compatibility	-.1554	-.0219	-.0340	0.0910	0.0320	-.1729	0.1092	-.1276	-.0204	0.1133	0.1908	1.0000

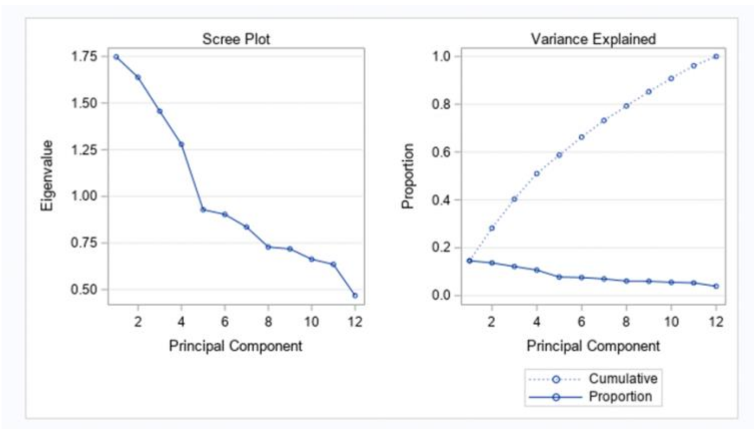
To perform clustering we will consider only the eigenvalues that exceed the average expected value of of the eigenvalues which is 1, this is known as the Kaiser-Guttman test. As a result, in our study Prin1-Prin4 are essential part of variability, while Prin5-Prin12 can be considered as noise.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.74763872	0.10945950	0.1456	0.1456
2	1.63817922	0.18158969	0.1365	0.2822
3	1.45658953	0.17791724	0.1214	0.4035
4	1.27867229	0.35028217	0.1066	0.5101
5	0.92839012	0.02525734	0.0774	0.5875
6	0.90313278	0.06762303	0.0753	0.6627
7	0.83550975	0.10718707	0.0696	0.7323
8	0.72832267	0.01021259	0.0607	0.7930
9	0.71811008	0.05554840	0.0598	0.8529
10	0.66256169	0.02754497	0.0552	0.9081
11	0.63501671	0.16714026	0.0529	0.9610
12	0.46787645		0.0390	1.0000

From the eigenvectors table we can easily observe the weights of variables in each component. For e.g., in Prin1 the new_1, new_8, new_10 and new_11 have the highest weights ans so are the most important variable in the first PC.

Eigenvectors													
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
new_1	wireless charging	-.087447	0.325314	0.431130	-.316333	-.201184	-.229916	0.063088	-.139693	0.092419	0.386320	0.569403	0.041824
new_2	finger sensor	-.155961	0.469669	-.041652	-.024471	0.158994	0.504889	0.221437	0.446339	0.321956	-.277145	0.208128	-.022933
new_3	multi cameras	0.204870	0.142545	0.391315	0.351815	0.086058	0.117937	-.614300	-.280695	0.220492	-.323233	0.099448	0.142662
new_4	OLED display	0.333701	0.354384	0.065967	0.333517	0.273535	0.197417	0.218563	-.080757	-.129801	0.562381	-.264340	0.276386
new_5	simple interface	0.277849	0.218425	-.121948	-.368458	0.564171	-.277334	-.099769	0.064085	-.436021	-.281392	0.171942	0.127802
new_6	different colors	-.140516	-.452972	0.286812	-.198048	0.132003	0.310471	-.295726	0.452228	-.147543	0.251796	0.049455	0.400239
new_7	fits in pocket	0.057587	-.195573	-.545500	0.174454	0.290491	-.175917	-.176888	0.026347	0.472842	0.295055	0.404288	0.095022
new_8	gaming	-.084553	-.150441	0.219796	0.594108	0.000875	-.370254	0.373617	0.280213	-.175506	-.228350	0.281493	0.227071
new_9	elegant design	0.166396	-.434370	0.165199	-.066780	0.294726	0.404661	0.415106	-.453254	0.000935	-.100951	0.301254	-.150242
new_10	brand	0.484637	-.118510	0.100145	-.305846	-.176510	-.180831	0.261779	0.085949	0.480989	-.192002	-.226569	0.433811
new_11	face ID	0.575435	-.097154	0.180946	0.072791	-.055900	-.032071	-.092962	0.440894	0.000966	0.139129	0.084267	-.622699
new_12	compatibility	0.337052	0.041707	-.376615	0.051428	-.557472	0.317053	-.063886	-.030189	-.353158	-.070706	0.358904	0.255637

Recall that for a principal component analysis (PCA) of *p* variables, a goal is to represent most of the variation in the data by using *k* new variables, where hopefully *k* is much smaller than *p*. Thus PCA is known as a *dimension-reduction algorithm*.



Moreover, if the scree plot contains an "elbow" (a sharp change in the slopes of adjacent line segments), that location might indicate a good number of principal components (PCs) to retain.

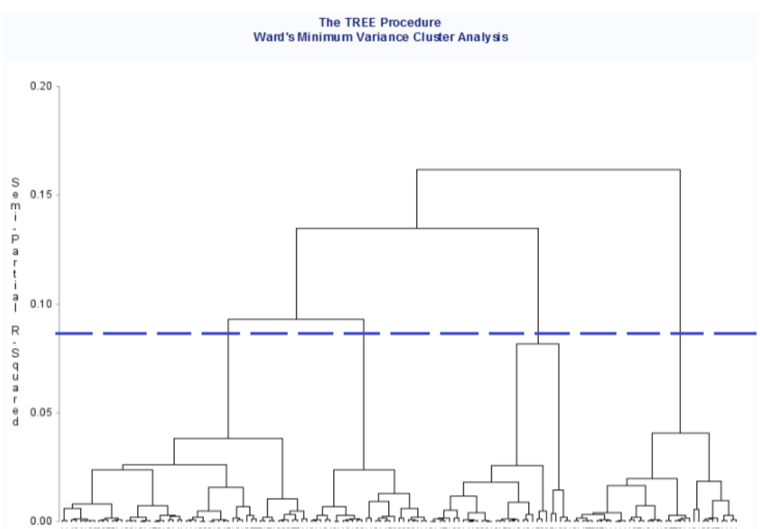
For this example, the scree plot shows a large change in slopes at the fourth eigenvalue. From the graph of the cumulative proportions, you can see that the first four PCs explain 51% of the variance in the data.

4. Clustering

4.1 Cluster identification and Dendrogram

Using the first four principal components, we applied the CLUSTER and TREE procedures with Ward as the method to plot a Dendrogram in order to determine the number of clusters we are going to use for the analysis:

```
proc cluster data=project.coord_new method=ward outtree=project.tree_new;
 var prin1-prin4;*selected using eigen value structured analysis;
 id id;
 run;
proc tree; run;
```



By analyzing the Dendrogram, we determine to use four clusters that are represented by the four groups of data under the blue dotted line. Afterwards, we created a frequency table of the clusters:

CLUSTER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	31	21.09	31	21.09
2	54	36.73	85	57.82
3	26	17.69	111	75.51
4	36	24.49	147	100.00

With the Dendrogram and the table, we observed that the four clusters are different between them, and they also have a similar amount of data, so we proceeded to do the cluster analysis in order to be able to determine which variables are statistically significant from to describe them.

4.2 Cluster Analysis

4.2.1 Gender

In order to determine if the gender explains the difference between the clusters, we applied a chi-square test by using the following code:

```
proc freq data=project.New_survey_new_2;
table gender*cluster / expected chisq;
run;
```

Giving the output:

Frequency / Expected / Percent / Row Pct / Col Pct

Table of Gender by CLUSTER

Gender(Gender)	CLUSTER 1	2	3	4	Total
Female	13 17.925 8.84 15.29 41.94	34 31.224 23.13 40.00 62.96	15 15.034 10.20 17.65 57.69	23 20.816 15.65 27.06 63.89	85 57.82
Male	17 12.864 11.56 27.87 54.84	20 22.408 13.61 32.79 37.04	11 10.789 7.48 18.03 42.31	13 14.939 8.84 21.31 36.11	61 41.50
Prefer not to say	1 0.2109 0.68 100.00 3.23	0 0.3673 0.00 0.00 0.00	0 0.1769 0.00 0.00 0.00	0 0.2449 0.00 0.00 0.00	1 0.68
Total	31 21.09	54 36.73	26 17.69	36 24.49	147 100.00

Statistic	DF	Value	Prob
Chi-Square	6	7.4154	0.2841

As we can observe from the value of chi-square, we can conclude that, in general, the gender variable is not significant and thus does not explain the way in which the clusters are divided. However, by looking at the frequency and the expected values from the first table, it seemed like the first cluster is indeed significant, so we performed another chi-squared test between the cluster number 1 against the others:

Statistic	DF	Value	Prob
Chi-Square	2	7.1421	0.0281

With this, we can determine that, for this specific cluster, the variable is significant and, as a result, it highlights a characteristic of it.

4.2.2 Age

Then, in order to determine if the age explains the difference between the clusters, we applied a chi-square test by using the following code:

```
proc freq data=project.New_survey_new_1;
  table Age_group cluster;
  run;
```

Giving the output:

Frequency Expected Percent Row Pct Col Pct	Table of Age_group by CLUSTER					
	Age_group(Age group)	CLUSTER				
		1	2	3	4	Total
	15-19	7 2.7415 4.76 53.85 22.58	1 4.7755 0.68 7.69 1.85	3 2.2993 2.04 23.08 11.54	2 3.1837 1.36 15.38 5.56	13 8.84
	20-24	4 3.585 2.72 23.53 12.90	8 6.2449 5.44 47.06 14.81	2 3.0068 1.36 11.76 7.69	3 4.1633 2.04 17.65 8.33	17 11.56
	25-29	8 13.497 5.44 12.50 25.81	34 23.51 23.13 53.13 62.96	11 11.32 7.48 17.19 42.31	11 15.673 7.48 17.19 30.56	64 43.54
	30-34	2 2.3197 1.36 18.18 6.45	4 4.0408 2.72 36.36 7.41	3 1.9456 2.04 27.27 11.54	2 2.6939 1.36 18.18 5.56	11 7.48
	35-40	3 1.898 2.04 33.33 9.68	1 3.3061 0.68 11.11 1.85	1 1.5918 0.68 11.11 3.85	4 2.2041 2.72 44.44 11.11	9 6.12
	Over 40	7 6.9592 4.76 21.21 22.58	6 12.122 4.08 18.18 11.11	6 5.8367 4.08 18.18 23.08	14 8.0816 9.52 42.42 38.89	33 22.45
	Total	31 21.09	54 36.73	26 17.69	36 24.49	147 100.00

Statistic	DF	Value	Prob
Chi-Square	15	31.9360	0.0066

As we can observe from the value of chi-square, we can conclude that, in general, the age variable is significant and thus it explains how the clusters are divided. Based on the general significance, by looking at the frequency and the expected values from the first table, we decided to perform another chi-squared test between each individual cluster and the others one by one:

Cluster number 1:

Statistic	DF	Value	Prob
Chi-Square	5	12.1474	0.0328

Cluster number 2:

Statistic	DF	Value	Prob
Chi-Square	5	20.3266	0.0011

With this, we can determine that, for clusters number 1 and 2, the variable age is significant and, as a result, it highlights a characteristic of these two clusters.

4.2.3 Income level

As with the other variables, we applied a chi-square test to the income level by using the following code:

```
proc freq data=project.New_survey_new_1;
table Average_annual_salary__euros_ cluster;
run;
```

Giving the output:

Table of Average_annual_salary__euros_ by CLUSTER

Average_annual_salary__euros_(Average annual salary (euros))	Frequency / Expected / Percent / Row Pct / Col Pct	CLUSTER 1	2	3	4	Total
0-15,000	Frequency	7	13	9	5	34
	Expected	7.1701	12.49	6.0136	8.3265	
	Percent	4.76	8.84	6.12	3.40	23.13
	Row Pct	20.59	38.24	26.47	14.71	
	Col Pct	22.58	24.07	34.62	13.89	
15,000-28,000	Frequency	5	12	4	5	26
	Expected	5.483	9.551	4.5986	6.3673	
	Percent	3.40	8.16	2.72	3.40	17.69
	Row Pct	19.23	46.15	15.38	19.23	
	Col Pct	16.13	22.22	15.38	13.89	
28,000-35,000	Frequency	2	4	0	7	13
	Expected	2.7415	4.7755	2.2993	3.1837	
	Percent	1.36	2.72	0.00	4.76	8.84
	Row Pct	15.38	30.77	0.00	53.85	
	Col Pct	6.45	7.41	0.00	19.44	
35,000 - 50,000	Frequency	3	3	1	6	13
	Expected	2.7415	4.7755	2.2993	3.1837	
	Percent	2.04	2.04	0.68	4.08	8.84
	Row Pct	23.08	23.08	7.69	46.15	
	Col Pct	9.68	5.56	3.85	16.67	
I don't have a salary yet.	Frequency	11	12	10	12	45
	Expected	9.4898	16.531	7.9592	11.02	
	Percent	7.48	8.16	6.80	8.16	30.61
	Row Pct	24.44	26.67	22.22	26.67	
	Col Pct	35.48	22.22	38.46	33.33	
More than 50,000	Frequency	3	10	2	1	16
	Expected	3.3741	5.8776	2.8299	3.9184	
	Percent	2.04	6.80	1.36	0.68	10.88
	Row Pct	18.75	62.50	12.50	6.25	
	Col Pct	9.68	18.52	7.69	2.78	
Total		31	54	26	36	147

Statistic	DF	Value	Prob
Chi-Square	15	22.4319	0.0970

As we can observe from the value of chi-square, we can conclude that, in general, the income level variable is not significant and thus does not explain the way in which the clusters are divided. However, by looking at the frequency and the expected values from the first table, it seemed like the second cluster is indeed significant, so we performed another chi-squared test between the cluster number 2 against the others:

Statistic	DF	Value	Prob
Chi-Square	5	8.8011	0.1173

In this case, we can determine that the variable income level is not significant.

4.2.4 Country

Afterwards, in order to determine if the country explains the difference between the clusters, we applied a chi-square test by using the following code as before:

```
proc freq data=project.New_survey_new_1;
 table Country_2 cluster;
 run;
```

Giving the output:

Table of Country_2 by CLUSTER						
		CLUSTER				
Country_2(Country 2)	Frequency / Expected / Percent / Row Pct / Col Pct	1	2	3	4	Total
Ecuador	Frequency	5	26	9	9	49
	Expected	10.333	18	8.6667	12	
	Percent	3.40	17.69	6.12	6.12	33.33
	Row Pct	10.20	53.06	18.37	18.37	
	Col Pct	16.13	48.15	34.62	25.00	
Italy	Frequency	21	15	11	22	69
	Expected	14.551	25.347	12.204	16.898	
	Percent	14.29	10.20	7.48	14.97	46.94
	Row Pct	30.43	21.74	15.94	31.88	
	Col Pct	67.74	27.78	42.31	61.11	
Other	Frequency	5	13	6	5	29
	Expected	6.1156	10.653	5.1293	7.102	
	Percent	3.40	8.84	4.08	3.40	19.73
	Row Pct	17.24	44.83	20.69	17.24	
	Col Pct	16.13	24.07	23.08	13.89	
Total	Frequency	31	54	26	36	147
	Percent	21.09	36.73	17.69	24.49	100.00

Statistic	DF	Value	Prob
Chi-Square	6	17.3028	0.0082

As we can observe from the value of chi-square, we can conclude that, in general, the country variable is significant and thus it explains how the clusters are divided. Based on the general significance, by looking at the frequency and the expected values from the first table, we decided to perform another chi-squared test between each individual cluster and the others one by one:

Cluster number 1:

Statistic	DF	Value	Prob
Chi-Square	2	7.3682	0.0251

Cluster number 2:

Statistic	DF	Value	Prob
Chi-Square	2	13.1136	0.0014

With this, we can determine that, for clusters number 1 and 2, the variable country is significant and, as a result, it highlights a characteristic of these two clusters.

4.2.5 Education Level

As a first step, we look at how the different levels of education are distributed across the different clusters and the expected chi-square value. We do so to understand how education level is related to the 4 clusters.

As a result, we run the following code:

```
proc freq data=project.New_survey_new_1;
table Education_level*cluster / expected chisq;
run;
```

The code above gives us the following output:

Frequency
Expected
Percent
Row Pct
Col Pct

Table of Education_level by CLUSTER					
	CLUSTER				
Education_level(Education level)	1	2	3	4	Total
Phd	0 0.2109 0.00 0.00 0.00	1 0.3673 0.68 100.00 1.85	0 0.1769 0.00 0.00 0.00	0 0.2449 0.00 0.00 0.00	1 0.68
Postgraduate	16 18.136 10.88 18.60 51.61	34 31.592 23.13 39.53 62.96	15 15.211 10.20 17.44 57.69	21 21.061 14.29 24.42 58.33	86 58.50
School	6 4.4286 4.08 28.57 19.35	7 7.7143 4.76 33.33 12.96	5 3.7143 3.40 23.81 19.23	3 5.1429 2.04 14.29 8.33	21 14.29
Undergraduate	9 8.2245 6.12 23.08 29.03	12 14.327 8.16 30.77 22.22	6 6.898 4.08 15.38 23.08	12 9.551 8.16 30.77 33.33	39 26.53
Total	31 21.09	54 36.73	26 17.69	36 24.49	147 100.00

Statistic	DF	Value	Prob
Chi-Square	9	5.3179	0.8058

As we can observe from the output table, by looking at the probability value of chi-square, we can conclude that the education level variable is not significant and thus does not explain the way in which the clusters are divided.

In fact, in each cluster we can observe that the frequency of each variable is quite similar to its expected variable. Therefore, we can conclude that education level is not a dependent variable for our clusters.

4.2.6 Working status

In this case instead, we want to understand how the “working status” variable is related to our clusters. Similarly to the previous examples, we run the following code:

```
proc freq data=project.New_survey_new_1;
```

```
table Working_status*cluster / expected chisq;
run;
```

Which outputs:

Frequency Expected Percent Row Pct Col Pct	Table of Working_status by CLUSTER				
Working_status(Working status)	CLUSTER 1	2	3	4	Total
Full-time worker	18 18.769 12.24 20.22 58.06	37 32.694 25.17 41.57 68.52	12 15.741 8.16 13.48 46.15	22 21.796 14.97 24.72 61.11	89 60.54
Other	1 1.6871 0.68 12.50 3.23	2 2.9388 1.36 25.00 3.70	2 1.415 1.36 25.00 7.69	3 1.9592 2.04 37.50 8.33	8 5.44
Part-time worker	2 1.898 1.36 22.22 6.45	4 3.3061 2.72 44.44 7.41	1 1.5918 0.68 11.11 3.85	2 2.2041 1.36 22.22 5.56	9 6.12
Retired	0 0.4218 0.00 0.00 0.00	0 0.7347 0.00 0.00 0.00	1 0.3537 0.68 50.00 3.85	1 0.4898 0.68 50.00 2.78	2 1.36
Student	10 8.2245 6.80 25.64 32.26	11 14.327 7.48 28.21 20.37	10 6.898 6.80 25.64 38.46	8 9.551 5.44 20.51 22.22	39 26.53
Total	31 21.09	54 36.73	26 17.69	36 24.49	147 100.00

Statistic	DF	Value	Prob
Chi-Square	12	8.9256	0.7093

As in the “Education level” analysis, the “working status” variable is not significant and does not help us to explain the clusters’ composition. Thus, we cannot consider it as a dependent variable.

4.2.7 Amount spent

Now, we want to understand how the “Amount spent” variable is related to our clusters.
By running our code, we obtain:

Frequency
Expected
Percent
Row Pct
Col Pct

Table of Amount_spent_on_a_mobile_phone__ by CLUSTER

Amount_spent_on_a_mobile_phone__(Amount spent on a mobile phone (euros))	CLUSTER 1	2	3	4	Total
200-400	23 11.599 15.65 41.82 74.19	14 20.204 9.52 25.45 25.93	11 9.7279 7.48 20.00 42.31	7 13.469 4.76 12.73 19.44	55 37.41
400-600	5 4.4286 3.40 23.81 16.13	7 7.7143 4.76 33.33 12.96	0 3.7143 0.00 0.00 0.00	9 5.1429 6.12 42.86 25.00	21 14.29
600-800	2 6.1156 1.36 6.90 6.45	17 10.653 11.56 58.62 31.48	4 5.1293 2.72 13.79 15.38	6 7.102 4.08 20.69 16.67	29 19.73
800-1,000	1 5.9048 0.68 3.57 3.23	11 10.286 7.48 39.29 20.37	5 4.9524 3.40 17.86 19.23	11 6.8571 7.48 39.29 30.56	28 19.05
More than 1,000	0 2.9524 0.00 0.00 0.00	5 5.1429 3.40 35.71 9.26	6 2.4762 4.08 42.86 23.08	3 3.4286 2.04 21.43 8.33	14 9.52
Total	31 21.09	54 36.73	26 17.69	36 24.49	147 100.00

Statistic	DF	Value	Prob
Chi-Square	12	44.7556	<.0001

The “Amount spent” variable is significant as we observe a strong enough difference between the observed and the expected values in all our clusters. For this reason, it is interesting to observe its relationship with each cluster:

First, we test the specific characterization given by “Amount spent” in cluster number 1. The variable is significant and, as a result, the variable is a dependent one:

Statistic	DF	Value	Prob
Chi-Square	4	26.7101	<.0001

This applies also to cluster number 2 even if at a lower extent since Chi-square is a bit lower and its probability higher:

Statistic	DF	Value	Prob
Chi-Square	4	9.1776	0.0568

The same applies for cluster number 3:

Statistic	DF	Value	Prob
Chi-Square	4	11.1093	0.0254

And cluster number 4:

Statistic	DF	Value	Prob
Chi-Square	4	11.5583	0.0210

In general, we can state that this variable is really good at explaining our clusters.

4.2.8 Brand

At this point, we want to understand how the "Brand" variable relates to our clusters. We start by running our initial code with the following output:

Frequency
Expected
Percent
Row Pct
Col Pct

Table of Which_brand_of_smartphone_did_yo by CLUSTER

	CLUSTER				
Which_brand_of_smartphone_did_yo(Which brand of smartphone did you purchase last time?)	1	2	3	4	Total
Apple	5 16.871 3.40 6.25 16.13	41 29.388 27.89 51.25 75.93	16 14.15 10.88 20.00 61.54	18 19.592 12.24 22.50 50.00	80 54.42
Google	1 0.2109 0.68 100.00 3.23	0 0.3673 0.00 0.00 0.00	0 0.1769 0.00 0.00 0.00	0 0.2449 0.00 0.00 0.00	1 0.68
Honor	1 0.2109 0.68 100.00 3.23	0 0.3673 0.00 0.00 0.00	0 0.1769 0.00 0.00 0.00	0 0.2449 0.00 0.00 0.00	1 0.68
Huawei	4 2.3197 2.72 36.36 12.90	2 4.0408 1.36 18.18 3.70	2 1.9456 1.36 18.18 7.69	3 2.6939 2.04 27.27 8.33	11 7.48
OPPO	2 0.6327 1.36 66.67 6.45	0 1.102 0.00 0.00 0.00	0 0.5306 0.00 0.00 0.00	1 0.7347 0.68 33.33 2.78	3 2.04
One plus	1 0.2109 0.68 100.00 3.23	0 0.3673 0.00 0.00 0.00	0 0.1769 0.00 0.00 0.00	0 0.2449 0.00 0.00 0.00	1 0.68
Pixel	0 0.2109 0.00 0.00 0.00	1 0.3673 0.68 100.00 1.85	0 0.1769 0.00 0.00 0.00	0 0.2449 0.00 0.00 0.00	1 0.68
Samsung	10 7.381 6.80 28.57 32.26	8 12.857 5.44 22.86 14.81	6 6.1905 4.08 17.14 23.08	11 8.5714 7.48 31.43 30.56	35 23.81
Sony	0 0.2109 0.00 0.00 0.00	0 0.3673 0.00 0.00 0.00	0 0.1769 0.00 0.00 0.00	1 0.2449 0.68 100.00 2.78	1 0.68
Xiaomi	7 2.7415 4.76 53.85 22.58	2 4.7755 1.36 15.38 3.70	2 2.2993 1.36 15.38 7.69	2 3.1837 1.36 15.38 5.56	13 8.84
Total	31 21.09	54 36.73	26 17.69	36 24.49	147 100.00

Statistic	DF	Value	Prob
Chi-Square	27	48.4769	0.0068

As we can notice, this variable has a high Chi-square value and a low probability, meaning that it is significant.
As in the section 4.2.7, we check for all the clusters.

The variable is significant for cluster 1:

Statistic	DF	Value	Prob
Chi-Square	9	37.1928	<.0001

And for cluster 2:

Statistic	DF	Value	Prob
Chi-Square	9	20.1188	0.0172

The variable however, is not significant for cluster 3:

Statistic	DF	Value	Prob
Chi-Square	9	2.0693	0.9903

And for cluster 4:

Statistic	DF	Value	Prob
Chi-Square	9	6.2189	0.7178

In general, we can state that this variable is useful to explain some of our clusters (1 and 2), but not all of them.

4.2.9 Use

Finally, we check how the "Use" variable relates to our clusters:

Table of What_is_the_primary_use_of_your_ by CLUSTER

What_is_the_primary_use_of_your_(What is the primary use of your smartphone?)	Frequency / Expected / Percent / Row Pct / Col Pct	CLUSTER 1	2	3	4	Total
Business	Frequency	3	2	2	7	14
	Expected	2.9524	5.1429	2.4762	3.4286	
	Percent	2.04	1.36	1.36	4.76	9.52
	Row Pct	21.43	14.29	14.29	50.00	
	Col Pct	9.68	3.70	7.69	19.44	
Business, Gaming	Frequency	1	0	0	0	1
	Expected	0.2109	0.3673	0.1769	0.2449	
	Percent	0.68	0.00	0.00	0.00	0.68
	Row Pct	100.00	0.00	0.00	0.00	
	Col Pct	3.23	0.00	0.00	0.00	
Business, Photography	Frequency	0	1	0	0	1
	Expected	0.2109	0.3673	0.1769	0.2449	
	Percent	0.00	0.68	0.00	0.00	0.68
	Row Pct	0.00	100.00	0.00	0.00	
	Col Pct	0.00	1.85	0.00	0.00	
Business, Social	Frequency	4	5	2	3	14
	Expected	2.9524	5.1429	2.4762	3.4286	
	Percent	2.72	3.40	1.36	2.04	9.52
	Row Pct	28.57	35.71	14.29	21.43	
	Col Pct	12.90	9.26	7.69	8.33	
Business, Social, Gaming, Photography	Frequency	0	0	2	0	2
	Expected	0.4218	0.7347	0.3537	0.4898	
	Percent	0.00	0.00	1.36	0.00	1.36
	Row Pct	0.00	0.00	100.00	0.00	
	Col Pct	0.00	0.00	7.69	0.00	
Business, Social, Photography	Frequency	1	5	2	6	14
	Expected	2.9524	5.1429	2.4762	3.4286	
	Percent	0.68	3.40	1.36	4.08	9.52
	Row Pct	7.14	35.71	14.29	42.86	
	Col Pct	3.23	9.26	7.69	16.67	
Gaming	Frequency	1	0	0	2	3
	Expected	0.6327	1.102	0.5306	0.7347	
	Percent	0.68	0.00	0.00	1.36	2.04
	Row Pct	33.33	0.00	0.00	66.67	
	Col Pct	3.23	0.00	0.00	5.56	
Photography	Frequency	0	0	1	1	2
	Expected	0.4218	0.7347	0.3537	0.4898	
	Percent	0.00	0.00	0.68	0.68	1.36
	Row Pct	0.00	0.00	50.00	50.00	
	Col Pct	0.00	0.00	3.85	2.78	
Social	Frequency	14	32	8	10	64
	Expected	13.497	23.51	11.32	15.673	
	Percent	9.52	21.77	5.44	6.80	43.54
	Row Pct	21.88	50.00	12.50	15.63	
	Col Pct	45.16	59.26	30.77	27.78	
Social, Gaming	Frequency	4	1	3	1	9
	Expected	1.898	3.3061	1.5918	2.2041	
	Percent	2.72	0.68	2.04	0.68	6.12
	Row Pct	44.44	11.11	33.33	11.11	
	Col Pct	12.90	1.85	11.54	2.78	
Social, Gaming, Photography	Frequency	0	0	1	1	2
	Expected	0.4218	0.7347	0.3537	0.4898	
	Percent	0.00	0.00	0.68	0.68	1.36
	Row Pct	0.00	0.00	50.00	50.00	
	Col Pct	0.00	0.00	3.85	2.78	
Social, Photography	Frequency	3	8	5	5	21
	Expected	4.4286	7.7143	3.7143	5.1429	
	Percent	2.04	5.44	3.40	3.40	14.29
	Row Pct	14.29	38.10	23.81	23.81	
	Col Pct	9.68	14.81	19.23	13.89	
Total	Frequency	31	54	26	36	147
	Percent	21.09	36.73	17.69	24.49	100.00

Statistic	DF	Value	Prob
Chi-Square	33	46.9757	0.0544

As we can observe, the variable has some significance and, as a result, we should check how it relates to each cluster.

The variable does not seem to be significant for cluster 1:

Statistic	DF	Value	Prob
Chi-Square	11	11.5491	0.3985

The variable seems to be more relevant for cluster 2, even though we cannot define it sufficiently significant:

Statistic	DF	Value	Prob
Chi-Square	11	17.9823	0.0820

The variable is not significant for cluster 3:

Statistic	DF	Value	Prob
Chi-Square	11	16.8214	0.1133

And for cluster 4:

Statistic	DF	Value	Prob
Chi-Square	11	16.7388	0.1158

In general, we can state that the variable explains our clusters to some extent since it is slightly significant. However, it is not significant enough to fully explain our clusters.

4.3 T-test

After clustering, we perfomend the t-test analysis of our core questions in order to interpret our 4 clusters.

The T-Test or Student's T-Test is any statistical hypothesis test in which the test statistics (t-statistics) follows a Student's t distribution if the null hypothesis is supported.

As a result, we run the following code to test for the equality of means for a two-sample (independent group) t-test.

```
data project.ttest_all;
merge project.cl_ttest_1 project.cl_ttest_2 project.cl_ttest_3 project.cl_ttest_4;
by variable;
run;
```

The **t-values** in the table below represents the ratio of the difference between the sample mean and the given number to the standard error of the mean. Since that the standard error of the mean measures the variability of the sample mean, the smaller the standard error of the mean, the more likely that our sample mean is close to the true population mean.

The **p-value** (prob) is the probability (computed using t-distribution) of observing a greater absolute value of t under the null hypothesis. If p-value is less than the pre-specified alpha level (in our study 0.05) we will conclude that mean is statistically significantly different from zero.

Finally, the **DF** value in the table below represents the degrees of freedom for the sample t-test.

	Variable	Method	Variances	tvalue_1	DF	prob_1	tvalue_2	prob_2	tvalue_3	prob_3	tvalue_4	prob_4
1	new_1	Satterthwaite	Unequal	-0.48	58.024	0.6317	-1.42	0.1581	-2.94	0.0054	4.48	<.0001
2	new_10	Satterthwaite	Unequal	-5.65	55.415	<.0001	4.64	<.0001	0.22	0.8253	-0.05	0.9627
3	new_11	Satterthwaite	Unequal	-5.46	59.32	<.0001	0.92	0.3585	5.09	<.0001	0.40	0.6892
4	new_12	Satterthwaite	Unequal	-2.41	51.585	0.0210	2.04	0.0430	0.51	0.6125	0.20	0.8417
5	new_2	Satterthwaite	Unequal	0.92	68.776	0.3613	-2.76	0.0069	-1.11	0.2734	4.38	<.0001
6	new_3	Satterthwaite	Unequal	-1.96	82.003	0.0572	-2.63	0.0097	3.69	0.0006	3.82	0.0003
7	new_4	Satterthwaite	Unequal	-2.98	79.072	0.0046	-2.87	0.0049	2.92	0.0060	4.91	<.0001
8	new_5	Satterthwaite	Unequal	-1.72	59.853	0.0934	1.59	0.1143	-1.96	0.0584	1.56	0.1238
9	new_6	Satterthwaite	Unequal	1.26	127.93	0.2136	1.42	0.1595	0.37	0.7115	-5.64	<.0001
10	new_7	Satterthwaite	Unequal	0.66	52	0.5098	1.11	0.2713	1.10	0.2775	-2.70	0.0094
11	new_8	Satterthwaite	Unequal	1.22	52.826	0.2287	-3.84	0.0002	3.65	0.0009	-0.73	0.4693
12	new_9	Satterthwaite	Unequal	-0.83	59.862	0.4115	2.99	0.0034	2.22	0.0323	-4.58	<.0001

In order to make this table more readable we associated each label with the corresponding variable, applying the following code:

```
data project.ttest_all_1; set project.ttest_all;
descr='              ';
if variable='new_1' then descr='wireless charging';
if variable='new_2' then descr='finger sensor';
if variable='new_3' then descr='multi cameras';
if variable='new_4' then descr='OLED display';
if variable='new_5' then descr='simple interface';
if variable='new_6' then descr='different colors';
if variable='new_7' then descr='fits in pocket';
if variable='new_8' then descr='gaming';
if variable='new_9' then descr='elegant design';
if variable='new_10' then descr='brand';
if variable='new_11' then descr='face ID';
if variable='new_12' then descr='compatibility';
run;
```

	Variable	Method	Variances	tvalue_1	DF	prob_1	tvalue_2	prob_2	tvalue_3	prob_3	tvalue_4	prob_4	descr
1	new_1	Satterthwaite	Unequal	-0.48	58.024	0.6317	-1.42	0.1581	-2.94	0.0054	4.48	<.0001	wireless char
2	new_10	Satterthwaite	Unequal	-5.65	55.415	<.0001	4.64	<.0001	0.22	0.8253	-0.05	0.9627	brand
3	new_11	Satterthwaite	Unequal	-5.46	59.32	<.0001	0.92	0.3585	5.09	<.0001	0.40	0.6892	face ID
4	new_12	Satterthwaite	Unequal	-2.41	51.585	0.0210	2.04	0.0430	0.51	0.6125	0.20	0.8417	compatibility
5	new_2	Satterthwaite	Unequal	0.92	68.776	0.3613	-2.76	0.0069	-1.11	0.2734	4.38	<.0001	finger sensor
6	new_3	Satterthwaite	Unequal	-1.96	82.003	0.0572	-2.63	0.0097	3.69	0.0006	3.82	0.0003	multi cameras
7	new_4	Satterthwaite	Unequal	-2.98	79.072	0.0046	-2.87	0.0049	2.92	0.0060	4.91	<.0001	OLED display
8	new_5	Satterthwaite	Unequal	-1.72	59.853	0.0934	1.59	0.1143	-1.96	0.0584	1.56	0.1238	simple interf
9	new_6	Satterthwaite	Unequal	1.26	127.93	0.2136	1.42	0.1595	0.37	0.7115	-5.64	<.0001	different col
10	new_7	Satterthwaite	Unequal	0.66	52	0.5098	1.11	0.2713	1.10	0.2775	-2.70	0.0094	fits in pocke
11	new_8	Satterthwaite	Unequal	1.22	52.826	0.2287	-3.84	0.0002	3.65	0.0009	-0.73	0.4693	gaming
12	new_9	Satterthwaite	Unequal	-0.83	59.862	0.4115	2.99	0.0034	2.22	0.0323	-4.58	<.0001	elegant desig

Now, we can print the content of our dataset using the code below:

```
proc print data=project.ttest_all_1;
var descr tvalue: prob:;
run;
```

The SAS System

Obs	descr	tvalue_1	tvalue_2	tvalue_3	tvalue_4	prob_1	prob_2	prob_3	prob_4
1	wireless char	-0.48	-1.42	-2.94	4.48	0.6317	0.1581	0.0054	<.0001
2	brand	-5.65	4.64	0.22	-0.05	<.0001	<.0001	0.8253	0.9627
3	face ID	-5.46	0.92	5.09	0.40	<.0001	0.3585	<.0001	0.6892
4	compatibility	-2.41	2.04	0.51	0.20	0.0210	0.0430	0.6125	0.8417
5	finger sensor	0.92	-2.76	-1.11	4.38	0.3613	0.0069	0.2734	<.0001
6	multi cameras	-1.96	-2.63	3.69	3.82	0.0572	0.0097	0.0006	0.0003
7	OLED display	-2.98	-2.87	2.92	4.91	0.0046	0.0049	0.0060	<.0001
8	simple interf	-1.72	1.59	-1.96	1.56	0.0934	0.1143	0.0584	0.1238
9	different col	1.26	1.42	0.37	-5.64	0.2136	0.1595	0.7115	<.0001
10	fits in pocke	0.66	1.11	1.10	-2.70	0.5098	0.2713	0.2775	0.0094
11	gaming	1.22	-3.84	3.65	-0.73	0.2287	0.0002	0.0009	0.4693
12	elegant desig	-0.83	2.99	2.22	-4.58	0.4115	0.0034	0.0323	<.0001

The table above can be divided into two main parts:

- The first four columns represents the t-values for each cluster. By looking at the values we are able to interpret the difference between each cluster. For example, the t-value for the cluster 4 regarding the OLED display question tells us that this feature is more important for this group than for the others.
- The last four columns represent the p-values. As we can observe the majority of these values are significant since we are using active variables to describe our clusters. However, there are also some of them that are not significant, this happens because we are in a "neutral" situation.

In the next section we will describe the differences between the clusters using both the categorical (chi-squared tests) and and the core questions (t-test).

5. Cluster Description

5.1 CLUSTER 1 - "Ragazzi"

The first cluster has 31 participants which represents the 21% of the whole sample and has the following characteristics:

- This cluster differenciates from the rest as it is composed mostly by men. This can be determined by observing the chi-square probability of 0.0281 for this cluster.
- The age is significant with a chi-square probability of 0.0328, showing that this cluster is composed by people between 15-19 years.
- By observing the chi-square probability of 0.0251, we determine that country is also significant for this cluster being composed mainly by Italians.
- This cluster differenciates from the rest as people spend between 200-400 euros in their smartphone. This can be determined by observing the chi-square probability of <0.0001 for this cluster.
- This group doesn´t care about the phone's brand or face ID feature as seen in the t-values of –5.65 and –5.49 respectively, both with a significance level of <0.0001.

5.2 CLUSTER 2 - “Ecuadorian Apple lovers”

This cluster is composed by 54 people that represents approximately 37% of the sample and has the following characteristics:

- This cluster differenciates from the rest because they prefer Apple’s smartphones. This can be determined by observing the chi-square probability of 0.0172 for this cluster.
- The age is significant with a chi-square probability of 0.0011, showing that this cluster is composed by people between 25-39 years.
- By observing the chi-square probability of 0.0014, we determine that country is also significant for this cluster being composed mainly by Ecuadorians.
- This group gives a lot of importance to the brand with a t-value of 4.64 (<0.0001) while they don´t care about the OLED display and the gaming features with t-values of –2.87 and –3.84 respectively, being both significant.

5.3 CLUSTER 3 - “Gamers”

This cluster is composed by 26 people that represents approximately 18% of the sample, 42% of the cluster is in the 25-29 age range and the cluster has the following characteristics:

- This cluster differenciates from the rest because they are the only cluster that has a strong preference for smartphones with a good gaming capacity.

- Members of this cluster do not come from a specific region, since they are almost equally distributed between Italy, Ecuador, and other countries.
- The amount spent is significant, with a chi-square probability of 0.0254, showing that 42% of the members spend between 200-400 euros, 16% spend between 600-800 euros, 19% spend between 800-1000 euros, and 23% spend more than 1000 euros
- For this cluster, Wireless charging is not important as can be observed from the negative t value –2.44. , being significant (prob. 0.0054)
- The most important features are: Face ID, with a t-value equal to 5.09 (<0.0001), Multi cameras with a t-value equal to 3.69 (0.0006), OLED display with a t-value equal to 2.92 (0.0060), Gaming with a t-value equal to 3.65 (0.0009), Elegant design with a t-value equal to 2.22 (0.0323)

5.4 CLUSTER 4 - "Geeks"

This cluster is composed by 36 people that represents approximately 24.5% of the sample, and 60% of the members come from Italy.

- This cluster differenciates from the rest because they are the only cluster that values the technical features of a smartphone.
- The amount spent is significant, with a chi-square probability of 0.0210, showing that people in this cluster are generally spreaded among all the amount ranges with the exception of the 1000+ range (which counts only 3 people).
- The must have features are: Wireless charging, with a t-value equal to 4.48 (<0.0001), Finger sensor with a t-value equal to 4.38 (<0.0001), Multi cameras with a t-value equal to 3.82 (0.0003), OLED display with a t-value equal to 4.91 (<0.0001) .
- These users do not value: Elegant design with a t-value equal to –4.58 (<0.0001), Different colors with a t-value equal to –5.64 (<0.0001), Fits in pocket with a t-value equal to –2.70 (0.0094)

6. Conclusions

Over the last decade, the utility of smartphones has increased at an exponential rate. In this research we focused on the study of mobile phone feature preferences. For the assessment and evaluation of smartphone features, it is necessary to address various population groups, for this reason we focused on 15 different countries found in 3 different continents.
The results indicate that the features considered important when buying a new smartphone depend on the type of cluster our respondents fall in. As a result, a T-test analysis is performed to identify the significant difference within the groups in relation to the usage of smartphone features.

After our clustering analysis we identified 4 main type of clusters:

- Cluster 1 “Ragazzi”, which differenciates from the rest as it is composed mostly by 15-19 years old men (mainly Italians).
- Cluster 2 “Ecuadorian Apple lovers”, which differenciates from the rest because they prefer Apple’s smartphones.
- Cluster 3 “Gamers”, which differenciates from the rest because they are the only cluster that has a strong preference for smartphones with a good gaming capacity.
- Cluster 4 “Geeks”, which differenciates from the rest because they are the only cluster that values the technical features of a smartphone.

From our results, we can observe that Apple is the preferred smartphone producer. However, we can also observe that there exists a “niche” that can occupied by other producers. For example, Apple does not specialize in gaming smartphones and, as a result, other producers might satisfy the needs of consumers requiring gaming dedicated smartphones as we can observe in cluster 3.

On the other hand, cluster 2 shows us that some brands are better than others in specific regions. This might result as a detterent for new investments of competing producers since customers already have a strong preference for a specific brand.

Cluster 1 instead, shows an area where brands can compete to offer a budget smartphone to young consumers. In fact, even though these customers do not give importance to “fancy” features such as Face ID, they still need a smartphone that is fast enough to carry out their social tasks.
This is also the reason why smartphone producers have been investing a lot of resources in making good low-budget smartphones over the last years.

Finally, cluster 4 shows us that some consumers require high-end smartphones, which offer all the latest features. Producers as a result, try to satisfy their needs by introducing new high performing smartphones every year. Of course, these consumers are also willing to spend a high amount of money for a smartphone with all the latest features.
One interesting fact that we can observe is that income level does not influence these consumers, since they are often driven by a strong passion or interest towards technology.