UNIVERSITÀ CA' FOSCARI DI VENEZIA

Department of Management

Master's Degree in Management

Master Thesis

Saving behaviours in the Fintech Era. The App Gimme5

Supervisor

Professor Ugo Rigoni

Graduand

Filippo Giuseppe Giusti

Matriculation: 861915

Academic year 2020/2021

Table of Contents

Introduction

In this thesis, we tried to find patterns in saving behaviors of people, analyzing data from a fintech App named Gimme5, which encourages people to save more using technology. The fintech App provided us a dataset containing different demographic and behavioral variables for almost fifty thousand of their users.

At the beginning of the thesis, in the first chapter, we presented the current state of the art explaining the reasons behind saving, analyzing, and commenting the main theories about saving behaviors, what drives saving and how these patterns can change among countries.

In the second chapter, we described the App Gimme5, and we performed a descriptive analysis to assess and understand the dataset. Concerning the App, we described the logic behind the functionality of saving and investing in Gimme5, describing how it is possible to save and where and for which reasons a user can invest. After that, we performed a descriptive analysis, which allows us to study the composition of the dataset from a demographic point of view (age, gender, bands of wealth, and residence) and assesses which users are relevant from a business point of view for Gimme5.

In the third and final chapter, we explained and used unsupervised statistical techniques to individuate groups of users with different sizes, behaviors, and demographic variables. First, we have applied PCA[1] analysis to find a low dimensional representation of the dataset and visually understand whether it is possible to identify clusters of "Active" users. After that, we used Elbow and Silhouette analysis to define the proper number of clusters and performed K-means and hierarchical clustering, enhancing and explaining our results. We tried a different number of clusters and unsupervised statistical techniques, but the major finding was detected using Silhouette analysis to individuate the most suitable number of clusters (5) for the dataset and the k-means clustering to individuate the most interesting patterns between groups. Our analysis allowed us to discover five groups of users, that give insights about the Gimme5 users, which are identifiable, sustainable, accessible, and actionable from a managerial point of view.

[1] Principal Component Analysis

To perform our analysis, we used a business intelligence tool (Qlik sense) and programming languages (R and Python) to aggregate and calculate new variables for our analysis. In particular, through the use of descriptive analysis, we were able to eliminate the users that are not relevant from a business point of view and to preserve, for statistical analysis, 20,000 users who are the most active in the platform and consequently more relevant for Gimme5. We individuated 8 variables and 20,000 users where we performed three unsupervised learning techniques (principal components analysis, K-means, and hierarchical clustering) to individuate a group of users with interesting characteristics for the business goals of Gimme5. After that, we underlined our findings using tables and graphs, which had the role to explain the differences among groups of users for the 8 variables taken into considerations.

We decided to write this thesis because the choice of how much to save is crucial for people. While standard theoretical frameworks suggest the optimal amount individuals should be save, based on their future earnings potential, their desired age of retirement, and consumption needs, a large majority of individuals do not save enough. Gimme5, through its App, is trying to help people to save more and is trying to do so by changing the logic of saving people from the standard behavior of "*gain, spend and save*" to "*gain, save and spend*" thanks to an App. On the other hand, Gimme5 can gather data about saving behaviors that were not possible to collect before. Consequently, Gimme5 allows us to perform an analysis to find patterns between saving behaviors and demographic variables and to contribute to the study of saving behaviors.

Saving behavior: from the theories to the European and Italian situation

1. Why do people save?

At the beginning of this thesis, I would like to study first the current state of the art analyzing why people save. Understanding the motives for which people save will allow us to analyze better the saving behaviors of households.

During the decades many theories have been discussed by economists and psychologists from all around the world. The stream of study can be grouped into four main directories. The formers are Keynesian-inspired, while the latters are more rooted in neo-classical principles. The first articulated theory of aggregate consumption responds to the so-called absolute income hypothesis (AIH) and is directly derived from a reading of General Theory. The second theory is known as the relative income hypothesis (RIH) and emphasizes how to aggregate consumption actually depends on the distribution of the same among classes of recipients with homogeneous lifestyles. Finally, we have the life-cycle hypothesis (LCH) tied to the natural processing of life in human-being and the permanent income hypothesis (PIH) related to the income seen as a constant installment derived from the wealth of an individual.

1.1 The absolute income hypothesis

The General Theory written by Keynes and published in 1936 can be represented from the four following statements[2]:

1. the marginal propensity to consume (MPC) is an established function of the available income Y;

[2] KEYNES J.C., *The General Theory of Employment, Interest, and Money,* Palgrave Macmillan, United Kingdom, *1936.*

2. the marginal propensity to consume (MPC) has a value greater than 0 but less than unity;
3. as income increases, the average propensity to consume (APC) decreases; with APC always greater than MPC;
4. as income increases, the value of MPC decreases.

In particular, Hypothesis two derives from the "*fundamental psychological law*" of Keynes. This states that individuals are willing to increase their consumption or savings as their income increases, but not to the extent of the increase in income itself. In fact, he argued that the satisfaction of the immediate basic needs of an individual and his family is usually a stronger motivation than the reasons that lead instead to accumulation, which acquire a real relevance when you have achieved a certain status or economic ease.

1.2 Relative income hypothesis

This hypothesis, which was advanced and developed primarily by James Duesenberry (1949), represents one of the first important contributions to the advancement of consumption theory.

Duesenberry's analysis is based on two assumptions[3]. First, at any given instant, utility functions are socially determined; in other words, they are interdependent across individuals. This can be evidenced by studying a cross-section of individuals at a given instant. Within each income group, a standard of living will be defined that is considered customary. Families with incomes lower than that considered "normal" will try to conform their consumption and savings to the latter, spending a higher fraction of their income so that the average propensity to consume and save of these families is higher than those of the norm, and vice versa: this is the so-called "demonstration effect". The second hypothesis is that utility functions are interdependent for the same individual at different times. For example, a family that experiences an increase in its income will, over time, learn to adjust its consumption and saving habits. Having learned to appreciate a higher

[3] DUESENBERRT J., *Income, Saving, and the Theory of Consumer Behavior*, Harvard University Press, United States, 1949.

standard of living, the family will not return to its original path of consumption, even if income were to return to its previous level. This is referred to in this case as the "harpoon" effect.

1.3 The lifecycle hypothesis

The life cycle hypothesis in consumption is a corpus of theories associated with the work of Franco Modigliani (1954) (1963) and his collaborators Albert Ando and Richard Brumberg. In accordance with these theories[4], the current consumption and savings of an individual depends on and is a fraction (in turn relative to tastes and preferences), of the current value of his vital resources, which are composed of personal wealth and earnings acquired during life (both current income and the expected future value of labor income). It is assumed that an individual maximizes his utility by maintaining stable, or with a little variation, the path of consumption over the course of his entire life. According to Modigliani, income and savings vary throughout an individual's life cycle. The life of the individual is divided into three phases. In the first and third stages, the individual does not work for exactly the opposite reasons, but clearly consumes the wealth of his or her own; in the second stage, the individual, while consuming, accumulates wealth because he or she is of working age. In the early years of his or her working life, the individual's income is relatively low and thus the individual typically connotes himself or herself as a net consumer or a net debtor (e.g., to finance the purchase of a home or consumer durables). Over the course of his or her working life, as income increases, the individual typically manages to accumulate useful assets to repay previously incurred debts, and to save in order to finance consumption for retirement age.

[4] MODIGLIANI F., BRUMBERG R., *Utility analysis and the consumption function: an interpretation of cross-section data*, Post-Keynesian economics., 1954, pp 388–436.
ANDO, MODIGLIANI F., *The 'life-cycle hypothesis of saving: aggregate implications and tests, "American Economic Review,* 1963, Vol. 53(1), pp. 55–84.

1.4 The permanent Income hypothesis

As noted above, in the early 1950s two similar yet distinct theories of consumption were proposed: the permanent income hypothesis and the life cycle hypothesis. Both of these theories can be seen as a response to the failure of the absolute income hypothesis to explain the empirical evidence. The permanent income hypothesis is attributable to Milton Friedman (1957) and his collaborators Rose Friedman, Dorothy Brady, and Margaret Reid. This hypothesis[5] is strictly a microeconomic theory of individual behavior. The analysis performed by Friedman is based on Fisher's (1907, 1930) theory of saving. The consumer plans consumption and savings, maximizing the long-run utility function, subject to the wealth constraint. When this constrained optimization problem is solved, current consumption depends on wealth and the interest rate. In contrast to the absolute income hypothesis in the permanent income hypothesis, a change in current income affects current consumption only if it alters wealth. Note that in the context of Friedman's analysis, wealth includes not only non-human wealth but also human wealth, where the latter can be defined as the present value of the sum of current and future labor income. Friedman introduces the concept of permanent income. It is defined as "the amount of consumption that the individual makes (or thinks he can make) while keeping wealth intact". Assuming an individual with infinite life, the permanent income can be interpreted as the constant installment of the income derived from the wealth of the individual.

1.5 Reasons to save

Reasoning on the motives for which people save will enable to give us a frame to the discussions and understand better the saving behaviors of households.

Most of the motives for which households save can be grouped into the three following categories[6]:

[5] FRIEDMAN M., *A theory of the Consumption Function,* Princeton University Press, United States, *1957*.
[6] HARIOKA Y.C., WATANABE W., *Why Do People Save? A Micro-Analysis of Motives for Household Saving in Japan,* The Economic Journal, 1997, Vol 107(442), pp. 537-552.

1. Life-cycle motives, defined as motives that arise from temporary imbalances between income and expenditures at various stages in one's life cycle, which in turn are due to differences in timing between income and expenditure streams. Examples include saving for one's leisure, marriage, and retirement expenses, one's consumer durables and housing purchases, and one's children's education and marriage expenses.
2. Precautionary motives, defined as motives arising from uncertainties concerning future income and/or expenditures. Examples include saving for income fluctuations, unemployment, illness, accidents, natural disasters, and longevity risk.
3. The bequest motive, which arises from the desire to leave assets behind to one's children and other heirs in the form of inter vivos transfers and/or bequests.

Life-cycle motives and precautionary motives are both consistent with the life-cycle model, whereas bequest motives can be consistent with either the life cycle or dynasty models, depending on the nature of the bequest motive. If bequests are motivated by intergenerational altruism, they are consistent with the dynasty model, whereas if they are unintended or accidental bequests arising from longevity risk or are motivated by selfish motives, they are consistent with the life-cycle model.

2. What are the drivers of saving behaviors?

We have analyzed the state of the art of saving behavior theories but are there situations in which we can artificially increase or decrease savings?

Analysis has shown that saving can increase depending on the type of goals, the number of goals, the number of reminders connected with a saving goal, and the time frame in which saving behavior is considered.

Starting from these main patterns we have tried to understand if saving behavior can be guided.

Reviews of Locke's (1968) theory of conscious goals which regulates behaviors and the goal-setting research of Latham (1975) indicate that performance is typically higher with difficult goals than with easy goals, as long as the difficult goals are accepted by the individual[7]. This implies that the clarification of the objective can increase savings and the probability to save respect "do your best" goal settings treatments. Therefore, presenting a single saving goal lends to greater savings intention and actual saving than presenting multiple savings goals. Multiple goals stimulate trade-off between goals and increase the probability that people will remain in a deliberative mindset and defer actions. On the other hand, a single goal evokes a stronger implementation intention[8]. This happens because goal pursuance is characterized by two stages: an initial stage with a deliberative mindset, in which people are uncertain about their goals and seek to define the desired outcome by considering the trade-off among goals, and a subsequent stage with an implemental mindset in which people have already established the goals they wish to pursue and are considering when, where, and how to attain them[9]. Subsequently, people are "ready to save" when they have clarified and fixed their objectives. Those are considerations to take into account also fixing the number of goals because consumers with a single goal can also help to stimulate an implementation intention compared with multiple goals. Multiple goals evoke trade-off consideration among

[7] IVANCEVICH, J., *Different Goal Setting Treatments and Their Effects on Performance and Job Satisfaction*, Academy of Management Journal, 1977, Vol.20(3), pp. 406–419.
[8] SOMAN D., ZHAO M., *The Fewer the Better: Number of Goals and Savings Behavior,* 2011, Journal of Marketing Research, Vol. 48, pp. 944-957.
[9] GOLLWITZER P.M., *Implementation Intentions: Strong Effects of Simple Plans,* 1999, American Psychologist, Vol. 54, pp. 493-503.

goals, which leave people in a deliberative mindset and hinder them from goal-related actions. Moreover, research in the area of mental accounting[10] clarifies the processes that consumers might use to make spending and saving decisions: consumers make decisions in the narrow context of specific product categories. This research explains that the probability of spending and saving is different for monies that are categorized into different mental accounts. For example, money budgeted for entrainment will more likely be spent on entertainment than on shopping, and money earmarked as saving will more likely be saved than money in a "spending account".

Another element to take into account is the temporal framing of savings, in fact, research[11] demonstrates the power of temporal reframing to boost participation in a recurring deposit saving program. A wide body of research has suggested that one barrier to future-oriented behavior is the tension that consumers feel between what they may want to do in the present versus what they think they should do for the future. Along these lines, framing savings contributions can be perceived as less "painful" from consumers. Therefore, a more temporally granular saving behavior may increase the likelihood that a consumer would be willing to make a present-day sacrifice for future gains. Tests conducted on savings apps show that reframing saving programs from monthly to daily have quadrupled the number of recurring savers and increased also the number of low-income savers. Additionally, other research[12] provides evidence that remainders messages increase commitment to save for consumers who recently opened savings accounts. Empirical tests conducted on three different banks have demonstrated that getting reminders increase the likelihood of meeting savings goals and increase saving amounts as well, compared to a no-reminder group.

Finally, the effects of goal setting on saving behavior have been studied also in the field of Fintech app. A study[13] in particular, demonstrate that individuals save

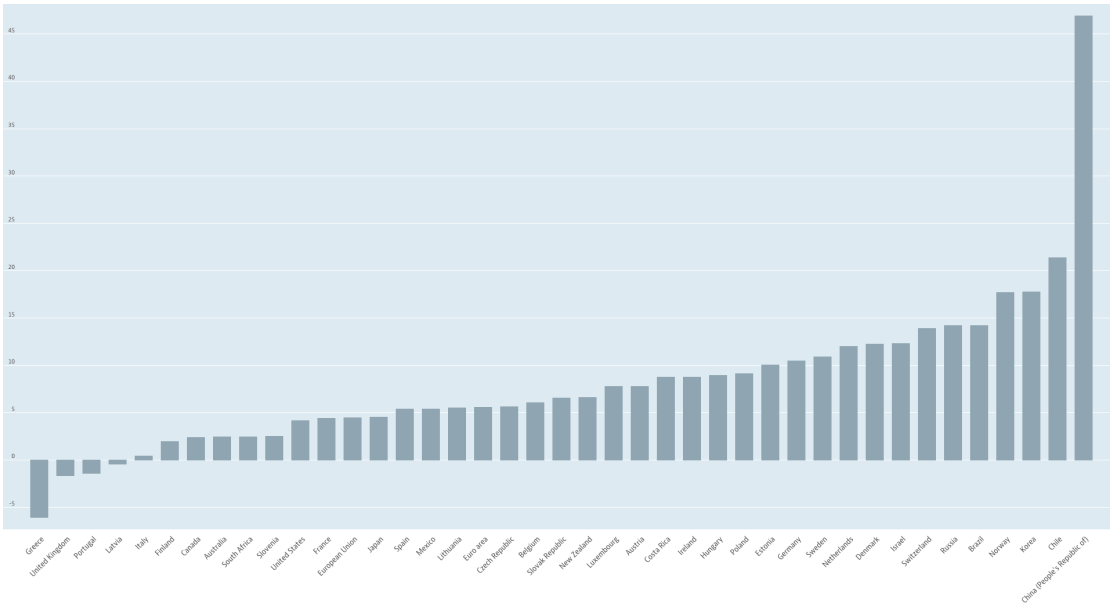
[10] THALER H.R., *Mental Accounting Matters,* Journal of Behavioral Decision Making, 1999, Vol.12, pp. 183-206.
[11] HERSHFIELD H., SHU S., BENARTZI S*., Temporal reframing and savings: A field experiment.,* 2019, SSRN Electronic Journal.
[12] KARLAN D., MCCONNELL M., ZINMAN J., *Getting to the top of mind: How reminders increase saving.* Management Science, 2016, Vol. 62(12), pp. 3393-3411.
[13] GARGANO A., ROSSI A., *There's an App for That: Goal setting and Saving in the FinTech Era,* 2020, SSRN Electronic Journal.

more for goals with shorter horizons and larger amounts. In 2020, a research team analyzed the introduction of a goal-setting feature in a Fintech app named "*Gimme5*" which allows investors to decompose their savings into different objectives, such as retirement, home, and car purchases, future travel, and many more. In addition, users could choose for each objective the horizon, the amount, and the mutual fund to invest the savings in. Results show that saving substantially increase in the first month after adopting goal setting and, over time, the additional saving decrease but stabilize always higher than the previous threshold. Moreover, tests demonstrate that short-term goals and small goals are the ones most likely to be achieved.

3. Saving across countries

Analyzing the saving from a geographical point of view, we can understand how much can differ the net saving rates, as a percentage of gross domestic product (GDP), among different nations.

Figure 1.1. Saving rates in percentage of GDP (2015)



Source: OECD

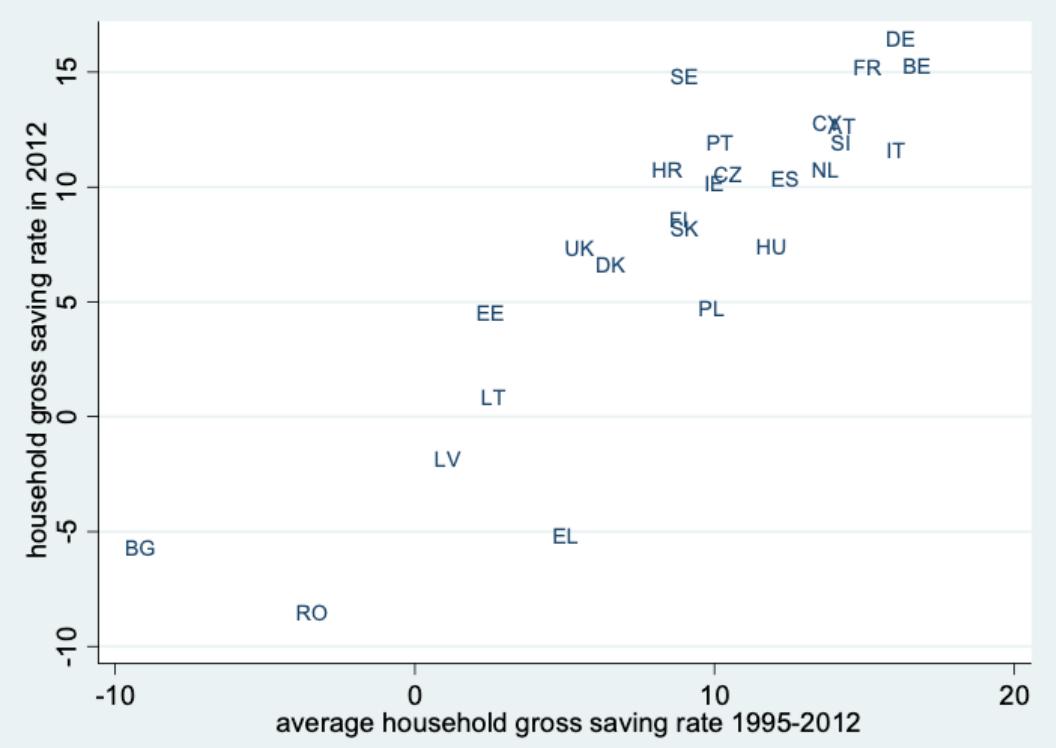
The graph shows that China is the country with the highest net saving in the world, the European Union average reaches almost 5% and Italy is less above 2%.

3.1 A focus on Europe

A Paper[14] has demonstrated that household saving rates differ significantly among EU countries and differences have proven to be persistent over time. In countries as Germany, France, and Belgium, households save a relatively large share of their disposable income. On the other hand, households in Romania and Bulgaria seem to spend often more than they earn, resulting in negative saving rates.

[14] ROCHER S., STIERLE M.*, Household saving rates in the EU: Why do they differ so much?,* 2015 Discussion Paper.

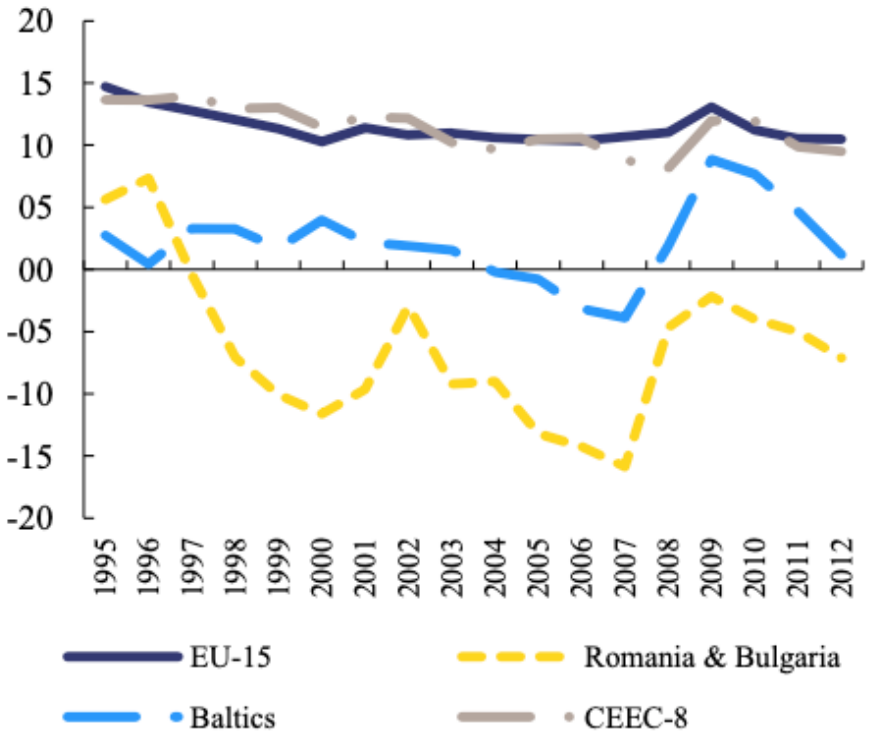
Figure 1.2 Persistence of households gross saving rates



Source: Eurostat

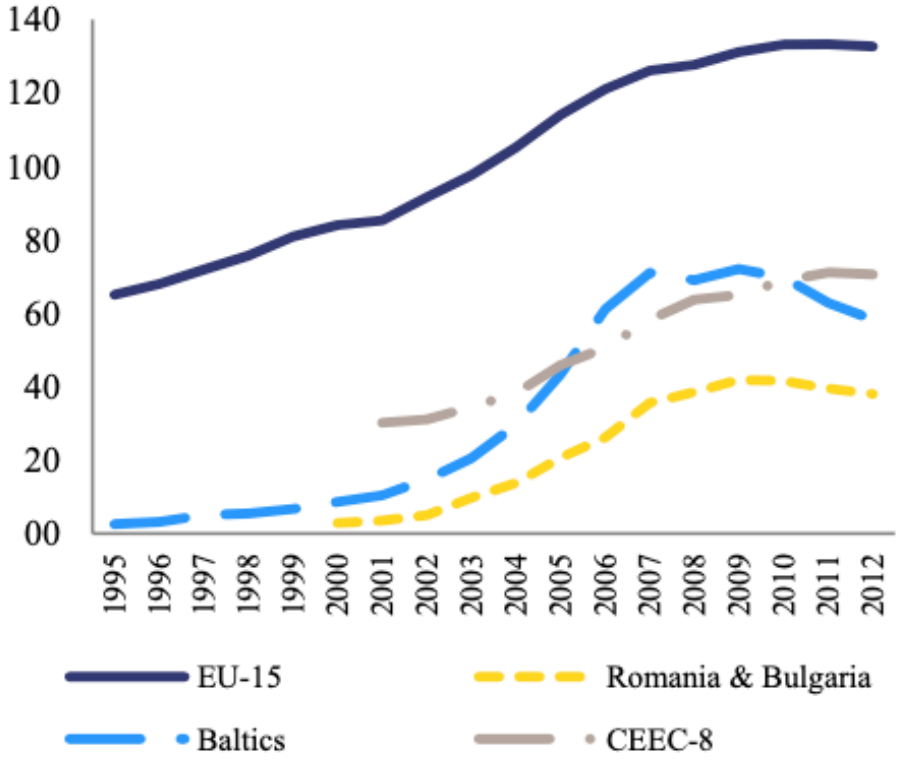
Due to data reliability and minimal international comparability, household saving rates, however, need to be read with caution. It seems that saving rates are in contrast to other economic variables. In Bulgaria (-11 percent) and Romania, for instance, saving rates have been negative over the last 15 years (-6 percent). This will mean that households spend dramatically and reliably more than they receive in these countries. It would be possible to doubt the sustainability of this situation. It does not completely reflect economic reality, however, as the debt-to-income ratios of households remain very low in these countries.

Figure 1.3 Household gross saving rate (in % of disposable income)



Source: Eurostat

Figure 1.4 Gross debt-to-income ratio of households (in %)



Source: Eurostat

3.1.1 Saving variables in the EU area

Several empirical studies have estimated the effect of various economic and demographic variables. Determinants of private savings can be divided into nine groups: Income, wealth, demographics, rates of return, uncertainty, fiscal policy, pension system, financial market performance, and international financial integration. These variables have different impacts on people's behaviors in the EU area.

3.1.2 Income and wealth

The saving rate is projected to increase with the so-called "income effect" level of revenue. The marginal tendency to save is expressed as the proportion of an additional euro in the disposable income that is saved. There is strong evidence that, when disposable income rises, the median tendency to save increases. The "wealth effect" predicts that wealthy people spend more, and all other things are equal, saving less of their money. Wealth will act as a buffer-stock and allows people to spend more of their money. Consequently, richer households tend to have lower saving rates.

In Europe, the per capita GDP amount has an insignificant effect on the household saving rate. Households save less than people in wealthy nations in poorer economies. Different factors can explain this result. Second, this may be empirical proof of the income impact that appears to explain much more of the saving rate differences across countries than within countries. Second, the amount of income is most likely to be associated with the fixed effect, catching unnoticed variations such as structural differences and other data problems. Third, any reverse causality can also be captured by the coefficient, as higher savings may lead to higher expenditure, an important source of economic growth and revenue. Instead, it is the change in real per capita GDP at the country level that explains shifts in household saving rates. Income growth appears to result in optimistic perceptions about the future. Therefore, rather than a convergence effect of household saving rates, we see a convergence of consumer spending followed by higher indebtedness.

3.1.3 Demography

According to the life-cycle hypothesis, young people tend to save less because of consumption smoothing over the lifetime, working-age people tend to save a lot, and elderly people tend to dis-save. It is also predicted that economies with high age dependence would have a lower aggregate household saving rate. Rising life expectancy, which offers anticipatory savings to brace an aging nation for longer retirement is projected to be favorably correlated with savings. However, analysis reveals that this coefficient has proved to be insignificant in Europe.

3.1.4 Uncertainty

Individuals prefer to invest more for precautionary purposes as they foresee bad times. One part of macroeconomic volatility may be the variance of inflation. The higher inflation volatility may also be assumed to associate favorably with household savings. In addition, since it promotes the possibility of being unemployed, the unemployment rate can be used as an indicator of income insecurity. In the Euro region, indicators of volatility have a positive effect on household savings. Low relative inflation rate and unemployment are boosting household savings.

3.1.5 Fiscal policy

In line with the Ricardian equivalence hypothesis, government deficits increase household savings. However, higher public debt levels in Europe appear to be associated with lower household savings. For precautionary reasons, we would expect households to save more in case of heavily indebted governments.

3.1.6 Financial market sophistication

The relation between the degree of maturity of the stock market and household saving is unclear. The growth of the financial system will, under

interesting circumstances, expand the possibilities for financial saving and promote household saving[15]. However, it also increases access to credit and eases households' liquidity constraints. Therefore, deeper financial markets may encourage smoother consumption, resulting in households borrowing more and saving less. In Europe[16], there is no evidence that the availability of liquidity in the economy can stimulate borrowing and thus reduce household savings, as others have found.

3.1.7 Financial Literacy

Financial is financial education, such as fundamental economics, statistics, and thinking abilities, paired with the ability to make financial choices using these skills. Analysis has shown that they make smarter investing and borrowing choices as individuals grow more financially literate, are more likely to save for retirement, and keep more varied investments in their balance sheet.

Figure 1.5 Financial Literacy around the world

Country/region	Number of countries	Literacy Score
EU	28	50
Non-EU advanced (excl. US)	8	58
US	1	57
China	1	28
Asia (excl. China)	12	32
Africa	35	33
Commonwealth of Independent States (CIS)	12	30
Latin America & Caribbean	19	29

Source: Standard & Poor's Global FinLit Survey.

Findings[17] about financial literacy in the EU echo similar findings from the US: lower-income individuals, women, and less-educated respondents rate lower than the rest of the population. Individuals below 25 years and above 70 years among

[15] PROCHNIAK M., WASIAK K., *The impact of the financial system on economic growth in the context of the global crisis: empirical evidence for the EU and OECD countries,* 2016, Empirica, Vol. 44, pp295-337.
[16] ROCHER S., STIERLE M., *Household saving rates in the EU: Why do they differ so much?,* Discussion Paper, 2015, pp 15-20.
[17] BATSAIKHAN U., DEMERTZIS M., *Financial literacy and inclusive growth in the European Union*, Bruegel-Policy Contribution, 2018.

55–65-year-old perform lowest and the right answer rate peaks. Saving, spending, and investment choices are important aspects of our lives, requiring an ever-growing degree of understanding of the risks and benefits that come with these decisions. In fact, higher financial literacy in individuals corresponds to better save and investment choices.

3.2 A focus on Italy

Narrowing our analysis, we will further investigate reasons for saving in Italy. This is possible thanks to the surveys carried out periodically by the main national bodies in order to understand the saving behavior of Italian families.

In 2018[18], more than 43 percent of "intentional" savers say they set aside resources to deal with unforeseen events; just under 20 percent save for old age; 21 percent do so for children; 14 percent for the home.

The data highlight a crucial point: for Italian families, savings have traditionally exercised (and still exercise) a fundamental insurance function. The other typical "virtue" of our country, which over time has had an insurance role, is strong social cohesion, particularly within the family. One figure is enough to give an idea of our country's solidarity capital: the latest census conducted by ISTAT on the sector shows that in 2015 more than 336,000 nonprofit institutions were operating in Italy, 11.6 percent more than in 2011, employing more than 5.5 million volunteers.

With the exception of car insurance policies, which have long been compulsory, the family and, more broadly, the network of relationships have always been among the main "insurers" of Italians against major risks, such as loss of employment or the onset of disabling illnesses. It is, therefore, no coincidence that, by international comparison, Italians are under-insured overall.

Italy has a particularly low ratio of non-life premium income to GDP: in 2016 the index reached 1.9 percent, stable compared to 2015 but down slightly from 2 percent in 2014. For comparison, the German figure was around 3.3 percent over

[18] RUSSO G., *Indagine sul Risparmio e sulle scelte Finanziarie degli Italiani,* 2018, Centro di ricerca e documentazione Luigi Einanudi e Intesa San Paolo, Torino.

the three years, while the French figure fluctuated between 3.2 and 3.4 percent. The Italian gap becomes even more evident when non-life premiums are excluded from total motor premiums: between 2014 and 2016, the ratio of non-motor premiums to GDP stopped at 0.9 percent in Italy, compared with 2.5 percent in Germany and 2.4 percent in France.

However, Italy is changing. As a natural consequence of the progressive aging of the population, the propensity to save is gradually declining, while continuing to support the accumulation of wealth. The structure of social relations is also changing, in the wake of demographic transformations: households are becoming smaller and smaller, while several generations tend to coexist for longer periods of time.

3.2.1 Reasons to save in Italy

In Italy, the main form of savings is generally precautionary[19]. This form of saving is always chosen by more than 40% of people and in some years has been the main reason for saving for Italians; it is particularly widespread among women, young people, and the elderly.

Saving for the home refers both to the setting aside of resources for the purchase and to the need to repay the mortgage, as well as to the need to renovate it; since the financial crisis, however, it has undergone a significant reduction that, albeit with various fluctuations, now places it at a level slightly more than half that of a decade ago. Men are more likely to save for their homes (18.2% compared to 7.2% of women); for reasons that are easy to understand, savings are particularly important between the ages of 25 and 44, while this motivation loses a lot of weight between the ages of 45 and 54 and undergoes a drastic reduction between the ages of 55 and older.

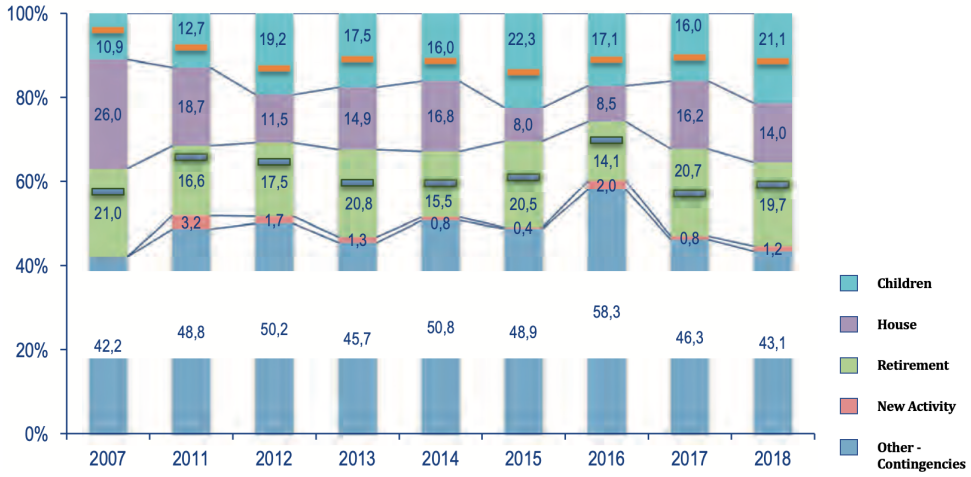
Saving for children, on the other hand, has grown over the years, almost doubling in the last decade. It includes education-related needs, resources to help children in their early independent years, and, of course, the inheritance motive,

[19] RUSSO G., *Indagine sul Risparmio e sulle scelte Finanziarie degli Italiani,* 2018, Centro di ricerca e documentazione Luigi Einanudi e Intesa San Paolo, Torino.

which generally absorbs about half of savings for children, with some variation from year to year. Both men and women set aside more or less equally for their children; they save more in the 45-64 age group, less over 65, and much less before 45.

Finally, saving for old age does not show a particular trend in the last decade: those for whom it is the main motivation oscillate around 20% of the sample. This saving includes both a generic reason, so to speak, for retirement, as well as health concern, and, in a decided minority proportion, the need for funds to move when one retires.

Figure 1.6 Savings Reasons in Italy



Sources: Intesa San Paolo

Saving in the FinTech Era

1. Gimme5: the digital moneybox

We've always been used to saving money by using a porcelain piggy bank, but in the Fintech era and thanks to digitization, more and more people decide to save through moneyboxes inside their smartphones. Currently, more and more Fintech companies are developing, and these companies can, with a simple App, make people save smarter. On the other hand, these Fintech companies allow us to gather more data about the saving and spending behaviors of people.

One, in particular, is Gimme5 which is one of the first fintech App in Europe to make people save smarter. Gimme5 is an initiative developed by AcomeA SGR, which is an asset management company independent of banking groups, led by professionals with long experience in the sector, 100,000 clients, and assets under management of approximately 2.5 billion euros. AcomeA was born in the 2010 from an idea of Alberto Foà, Giovanni Brambilla and Giordano Martinelli. These people had decades of experience in the financial sector and still now run the company receiving numerous international prizes for saving management. AcomeA has four main categories of funds: short-term monetary funds, bonds funds, flexible funds, and stocks funds.

Less than ten years ago, AcomeA launched Gimme5 App to keep up with the changing needs of its customers. In the beginning, the App served only as a digital piggy bank, but a few years later, they introduced goal-setting features that allowed investors to decompose their savings into different objectives, such as retirement, home, car purchases, future travel, etc... Besides, users could choose, for each objective, the horizon, the amount, and the mutual fund to invest the savings in. Now the App has sixty thousand users mainly located in Italy and every day helps people to save smarter.

1.1 Description of the App Gimme5

Gimme5 is an App-based company with the clear goal to help people to get more from their savings. To reach their goal, they have understood that people usually think first to spend and then save what's leftover. Therefore, the solution to help people to save more and reach their goals is to change the patterns of savings. Generally speaking, people save with a clear pattern:

Earn → Spend → Save.

On the other hand, AcomeA designed its App Gimme5 to change this pattern and help people to save more:

Earn → Save → Spend.

This will force people to settle before spending and consequently increase their savings.

Gimme5 business model is based on 4 pillars:

1. Save;
2. choose your speed;
3. get your objective;
4. get support.

With the pillar "save", users can take advantage of every opportunity to save, setting targets and activating automation to make simpler save. Speaking about the second pillar "choose your speed", users can choose the type of fund in which their monthly savings will be invested deciding the risk profile that suits best. The pillar "get your objectives" allows users to set their objectives and decide to save to reach the goal fixed, with a maximum of five savings goals. Finally, "get support" permits users to invite friends and family with the aim of support and help to reach their saving goals.

Observing more deeply, Gimme5 was first released in 2014 as a digital piggy bank that encouraged users to save small quantities of money. Individuals pay no activation costs and no recurring fees when they sign up. Investors will start investing as soon as they sign up for an account by depositing as much money as they want, anytime they want, just like a checking/savings account. Unlike checking and savings accounts, Gimme5 users can invest their money, no matter how small, in several investment funds with varying risk-return profiles. Individuals can spend their money in up to 13 funds with different management and performance fees[20] from a minimum of 0.60% to a maximum of 1.60% :

- AcomeA breve termine: the portfolio consists of government securities and bonds denominated in euros considered by Gimme5 with a low-to-medium level of risk and for people seeking a short-to-medium term investment (management fee.
- AcomeA Eurobbligazionario: the bond or monetary component of the portfolio is greater than 50% and it's considered by Gimme5 with a medium level of risk for people seeking a medium-term bond investment.
- AcomeA Performance: the portfolio is composed of more than 50% in bonds or monetary instruments and up to 15% in equities and it's considered by Gimme5 with a medium-high risk level for people seeking capital growth in the medium term.
- AcomeA patrimonio esente: flexible fund that implements an investment policy in line with the Regulations on Individual Savings Plans (PIR). The fund can invest in shares up to 40% of total assets and in bonds and monetary instruments up to 100% of total assets.
- AcomeA patrimonio prudente: The portfolio is composed of equity instruments up to 30% of the total and bond or monetary instruments up to 100% and it's considered by Gimme5 with a medium level of risk for people seeking growth of their capital in the medium term and with a medium level of risk.

[20] AcomeA official website, *Commissioni di gestione,* https://gimme5.acomea.it/costi/

- AcomeA patrimonio dinamico: the portfolio is composed of equity instruments up to 50% of the total and bonds or monetary instruments up to 100% and it's considered by Gimme5 with a medium to a high level of risk for those seeking growth of their capital in the medium to long-term for people seeking growth of their capital in the medium to long term.
- AcomeA patrimonio agressivo: the portfolio consists of equity instruments up to 100% of the total and bond or monetary instruments, up to 100% and, it's considered by Gimme5 with a high level of risk for people seeking capital growth in the long term.
- AcomeA globale: an international fund that invests at least 70% of its portfolio assets in stocks of companies around the world and, it's considered by Gimme5 with a high level of risk for people seeking long-term capital growth.
- AcomeA PMItalia ESG: a domestic fund that invests at least 70% of its portfolio assets in shares of Italian companies, it's considered by Gimme5 with a high level of risk for people seeking capital growth over the long term.
- AcomeA Europa: an international fund that invests at least 70% of the assets in the portfolio in shares of companies belonging to the European continent, it's considered by Gimme5 with a high-level of risk for people seeking long term capital growth.
- AcomeA America: an international fund that invests at least 70% of its portfolio assets in stocks of companies belonging to the American continent, it's considered by Gimmer5 with a low level of risk for people seeking long-term capital growth.
- AcomeA Asia Pacifico: the portfolio consists of at least 70% stocks of companies from the Asian and Oceanian continent, and it's considered by Gimme5 with a high-level of risk for people seeking long-term capital growth.
- AcomeA paesi emergent: the portfolio consists of at least 70% stocks of companies from emerging markets, and it's considered by Gimme5 with a high level of risk for people seeking long-term capital growth.

Moreover, the App added a goal-setting feature in October 2017, allowing users to set one or more saving goals. Gimme5's users are asked to choose an investment target from five broad categories while setting a goal:

- Hobby;
- travel;
- vehicle;
- home;
- general savings;
- other.

Besides, users must pick an investment horizon and a goal number and following the selection of the target, the investor selects the mutual fund into which she wants to position her funds. Thereafter three basic funds are shown in the app: "AcomeA breve termine", "AcomeA patrimonio dinamico", and "AcomeA patrimonio aggressivo". The aforementioned three funds, as we saw before, correspond to three main categories of risk: low, medium, and high. But Gimme5's users can also access the complete list of funds.

Once, decided a target and a fund, the investor will choose to deposit funds anytime she or he wishes, and the app, with a simple interface, tells users where they are on their way to achieving their goals.

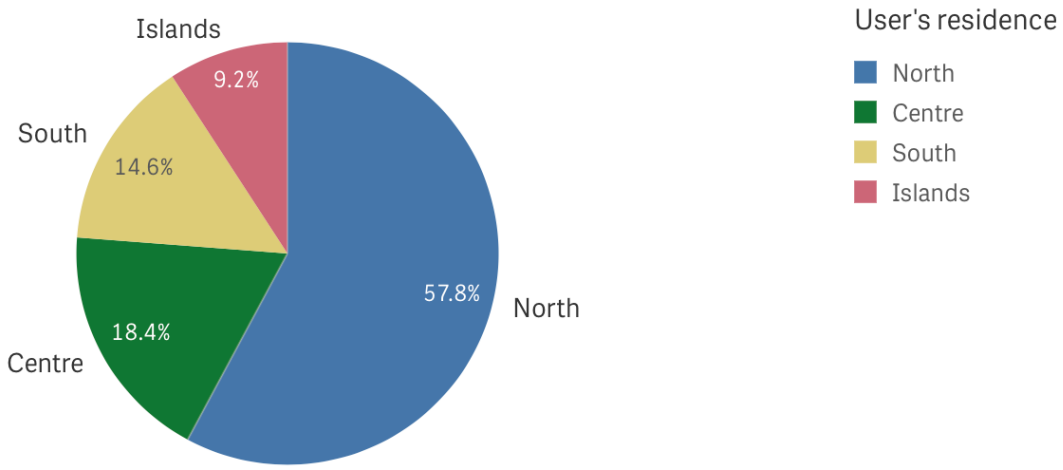
1.2 Gimme5's users

Thanks to the data shared by the company we were able to analyze the demographic and social characteristics of AcomeA users. We started our analysis from the raw data to understand the composition of the Gimme5 user's base studying variables such as user's residence, gender, age. After that, we create other variables with the data in our possession to deeper analyze our data and understand better the user's behavior. These new variables, which we will talk about specifically later, are user's bands of wealth and activity status. Users' bands of wealth represent the regional deposit bands published by the bank of Italy and activity status indicates whether a user is engaged with the platform or not.

1.2.1. User's residence

Studying the user's residence, we were able to assess the geographical distribution of Gimm5's members. First, as we can see from the pie chart below, we understood that more than half of Gimme5's users, precisely 57.8%, live in the north of Italy. After that we saw that other users are located for 18.4% in the center of Italy, 14.6% in the south of Italy, and only 9.2% live in Sicily or Sardinia.

Figure 2.1 User's residence

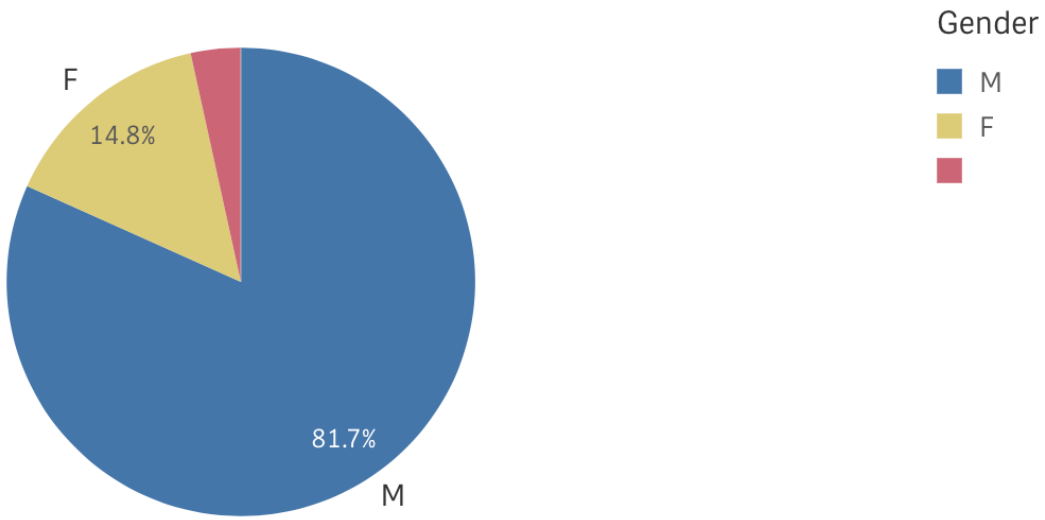


Source: Gimme5's dataset

1.2.2. User's Gender

Analyzing the sex variables, as you can see from the graph below, more than 81% of users are male and only 14.8% of users are female. There are also 3.5% of users that have decided to do not to indicate their sex. This analysis makes us realize that the vast majority of Gimme5's subscribers are male and only a small part is represented by female investors.

Figure 2.2 User's Gender

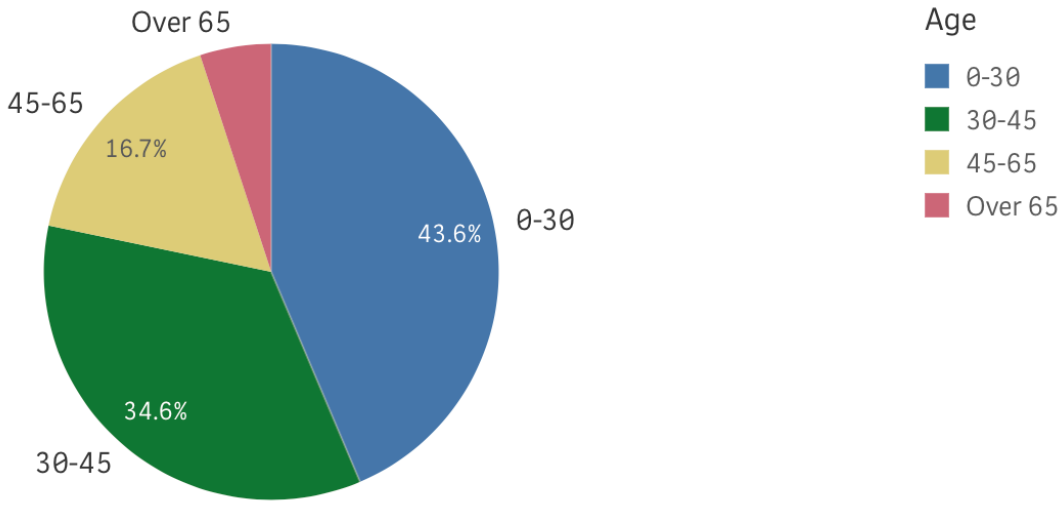


Source: Gimme5's dataset

1.2.3. User's Age

After the user's residence and sex, we decided to analyze the variable age. Studying the data, we have decided to divide the users into four age groups: 0-30, 30-45, 45-65, and over 65. As you can see from the pie chart below, we discover that the very large majority of users are in the age groups 0-30 (43.6%) or 30-45 (34.6%). On the other hand, the age group 45-65 and over 65 accounts respectively for the 16.7% and 5.1%.

Figure 2.3 User's Age



Source: Gimme5's dataset

1.2.4. User's bands of wealth

To deepen our analysis, we integrated some external data coming from the Bank of Italy. As can be seen from the table below, we took the number of deposits and the number of residents for each Italian region, and we created an index dividing deposits and habitants. This index can be considered as a first proxy of the financial wealth of the population. Segments have been chosen dividing first between two groups in correspondence of the index's average (22,98) and after that qualitatively identify the higher and the lower band containing the values that diverged from the central block. Doing this operation, we identified four bands of wealth: high, medium-high, medium-low, and low.

Figure 2.4 **User's bands of wealth**

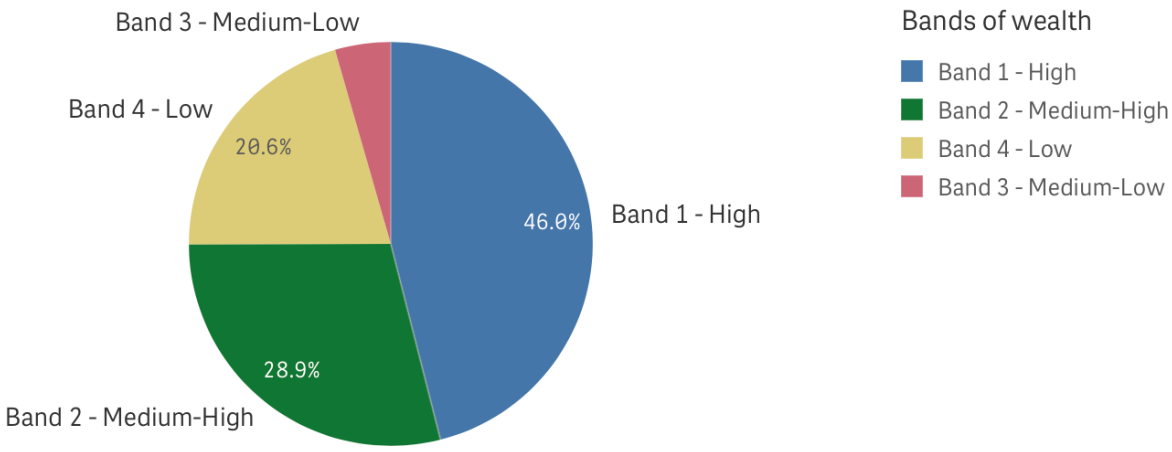
Regions	Deposits	Residents	Index
Lombardia	335.883.000	10.027.602	33,5
Emilia-Romagna	136.093.000	4.464.119	30,5
Valle d'Aosta	3.760.000	125.034	30,1
Lazio	164.762.000	5.755.700	28,6
Veneto	135.131.000	4.879.133	27,7
Piemonte	114.908.000	4.311.217	26,7
Friuli-Venezia Giulia	30.724.000	1.206.216	25,5
Liguria	38.521.000	1.524.826	25,3
Marche	37.457.000	1.512.672	24,8
Toscana	90.582.000	3.692.555	24,5
Trentino Alto Adige	26.403.000	1.078.069	24,5
Molise	6.382.000	300.516	21,2
Abruzzo	26.431.000	1.293.941	20,4
Basilicata	11.215.000	553.254	20,3
Umbria	17.555.000	870.165	20,2
Campania	98.406.000	5.712.143	17,2
Puglia	66.221.000	3.953.305	16,8
Sardegna	23.641.000	1.611.621	14,7
Calabria	26.924.000	1.894.110	14,2
Sicilia	63.085.000	4.875.290	12,9

Source: Bank of Italy

After identifying the bands of wealth and merging the data with Gimme5's dataset, we started to investigate which users belong to which bands of wealth.

What we discovered is that almost half (46%) of the subscribers belong to the high-income bracket, 28.6% of the members are in the medium-high income bracket, 20.6% in the low-income brackets, and only 0.5% belong to the medium-low-income bracket.

Figure 2.5 User's Band of wealth



Source: Gimme5's dataset

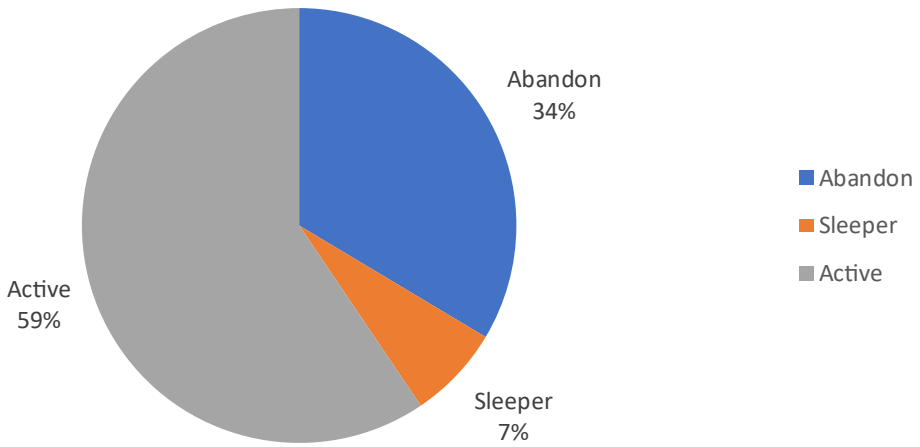
1.3 The Gimme5's user behaviors

In this paragraph, we explain how we decided to proceed with our analysis and the main decision we took to better examine our dataset. First of all, we decided to divide Gimme5's subscribers into two main sections: relevant and not relevant. The relevant members are those who have made more than one transaction on the platform. On the other hand, not relevant users are people who have performed only one transaction in the platform, and then they never used Gimme5 again. After making this division, we decided to exclude all the not relevant users and continue our analysis with the relevant members only.

Once excluded the not-relevant users, we focused on two variables: the number of transactions and the number of accesses. With these two variables, we executed two different analyses defining Gimme5's users active, sleeper, or abandoned depending on the number of transactions or the number of accesses they have performed.

Focusing on the number of transactions, we arbitrarily defined a Gimme5's user active if he or she has performed a transaction in the last 60 days, sleeper if a transaction occurred between 60 and 180 days ago, and abandon if a transaction last more than 180 days ago.

Figure 2.6 Transactions Status

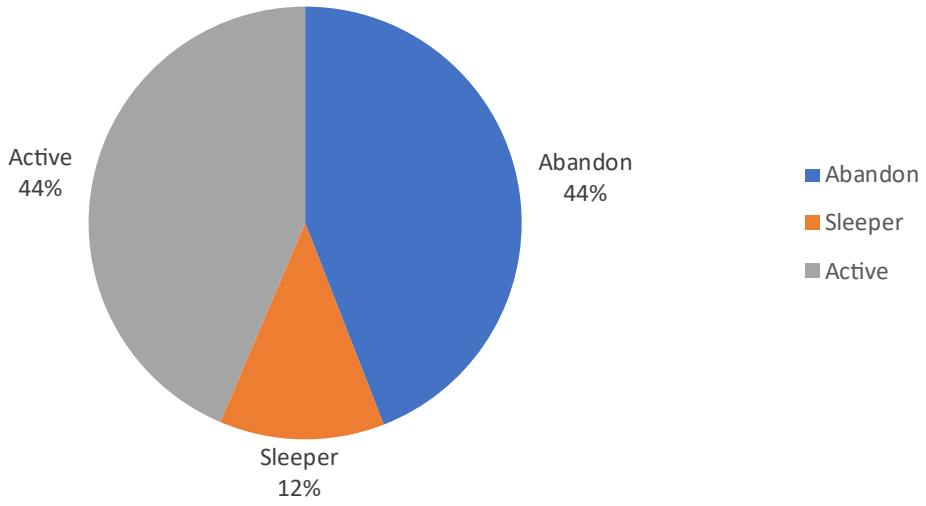


Source: Gimme5's dataset

Analyzing the graph, more than half of Gimme5's users (59%) are in an active status. Instead, the abandoned users are 34% of users and 7% of users are in the sleeper status.

On the other hand, we analyzed the number of accesses to define whether a user is active, sleeper, or abandon. In this second case, we defined a Gimme5's user as active if he or she has performed access in the last 60 days, sleeper if the last access occurred between 60 and 180 days ago, and abandon if the last access was carried out more than 180 days ago.

Figure 2.7 Accesses Status



Source: Gimme5's dataset

As we can see from the graphs, almost half of Gimme5's users (44%) are in abandon status. Instead, the active users have the same proportion (44%) of abandon users and the 12% of users are in the sleeper status.

Groups of users and patterns of behaviors in Gimme5

1. Methodology: Detecting cluster of users using Statistical techniques.

Statistical learning encompasses a wide range of methods for detecting patterns in data. Generally speaking, statistical learning is divided into two main streams: supervised and unsupervised learning**.** In general, supervised statistical learning entails creating a statistical model that predicts or estimates an outcome dependent on one or more inputs. This type of issue can arise in a variety of areas, including finance, medicine, astrophysics, and public policy. On the other hand, unsupervised statistical learning has inputs but no supervising output, therefore we can learn relationships and structure from such data.

1.1 Supervised vs Unsupervised Statistical Analysis

As said before, most statistical learning problems fall into two categories: supervised or unsupervised learning. Looking more deeply, supervised learning techniques assume that for each observation $x_i, \quad i = 1, \ldots, n$ there is a directly associated measurement y_i. Basically, we want to fit a model that relates the response to the predictors, with the aim to accurately predicting the response for future observations (prediction), or better understanding the relationship between the response and the predictors (inference)[21]. In this category are included many classical statistical learning methods such as linear regression and logistic regression, as well as more modern approaches such as generalize additive model (GAM), boosting, and support vector machine. Unsupervised learning, on the other side, explains the more complicated problem in which we observe a vector of

[21] JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., *An Introduction to statistical learning,* 2013, New York, Springer, page. 26-28.

measurements x_i but no corresponding response y_i for each observation $i = 1, \ldots, n$. Since there is no response variable a linear regression model cannot be fitted. In this context, we are in some sense working blind because we lack a response variable that can supervise our analysis. With unsupervised learning, we can seek to understand the relationships between the variables or between the observations. The most famous statistical learning tool of unsupervised learning is cluster analysis or clustering. The goal of cluster analysis is to ascertain, based on $x_i, \ldots, x_n$, whether the observations fall into relatively distinct groups. For example, we might observe several characteristics (variables) for potential consumers in a market segmentation analysis, such as zip code, family income, and shopping habits. We may assume that consumers are divided into various categories, such as high spenders and low spenders. Supervised research would be possible if knowledge about each customer's purchasing habits were available. This means that the company has to fix a variable not calculated as a result of other variables in the dataset that define a customer as high or low spender.

Therefore, if we don't have this information, we don't know if each potential buyer is a major spender or not, and consequently, we can't apply supervised learning to predict behaviors. In this case, we will attempt to classify consumers based on the variables available to distinguish distinct classes of possible customers. Identifying such groups may be interesting since the groups can vary in terms of any property of interest, such as spending habits.

1.2 Unsupervised Learning

For the nature of the Gimme5 dataset, we decided to apply unsupervised learning techniques to study different groups of users. Before starting the experiments, it is worth making an introduction to this complex subject to better explain the analysis performed below. In unsupervised learning we are not interested in prediction, because we do not have an associated response variable Y. Rather, the goal of these analyses is to discover interesting things in a dataset in which we have only a set of features $x_i, \ldots, x_n$, measured on n observations. Is there a way to visualize the data that is both insightful and interesting about Gimme5's users? Is it possible to find subgroups within the variables or the observations in

Gimme5's dataset? Are some questions we have tried to answer performing our analysis. In this chapter, we will focus on two particular types of unsupervised learning: principal components analysis, and clustering, a broad class of methods for discovering unknown subgroups in data.

1.2.1 Principal Components Analysis

Principal Components Analysis ("PCA") allows summarizing a dataset in a small number of variables that collectively explain most of the variability in the original set of variables. Suppose that we want to visualize n observations with a measurement of i features, $x_i, \ldots, x_n$. We can do this by examining two-dimensional scatterplots, each of which contains the n observations' measurement on two of the features. However, there are $\binom{i}{2} = i(i-1)/2$, scatterplots, this means that with $i = 10$ there are 45 plots. Clearly, it is necessary a better method to visualize the data. PCA is a tool to obtain a two-dimensional representation of data that captures most of the information and then plots the result on a Cartesian plane. Each of the n observations is said to exist in a i dimensional domain, but not all of these dimensions are equally interesting. The definition of interesting is determined by the degree that the observations vary in each dimension, and PCA finds a minimal number of measurements that are as interesting as possible.

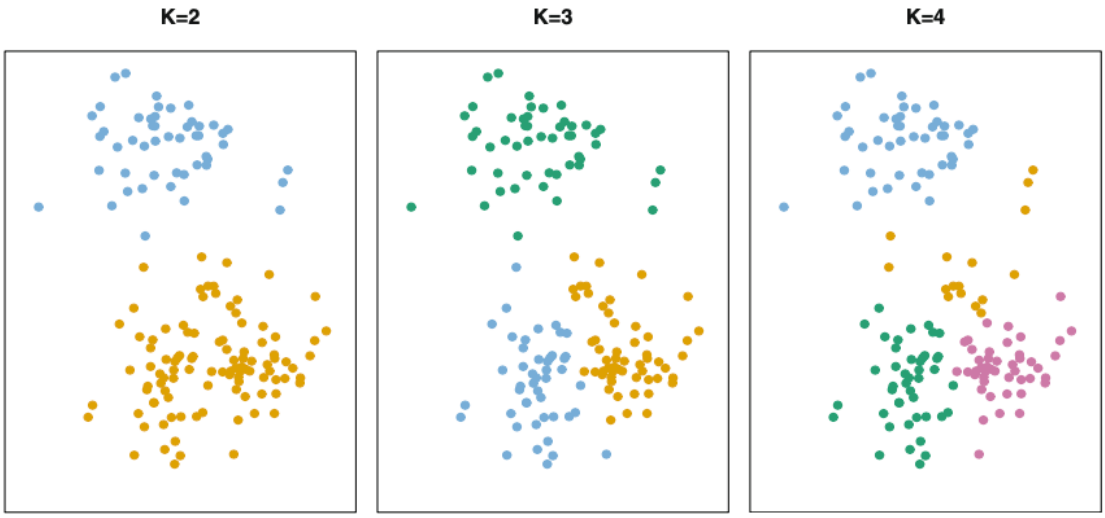
1.2.2 Clustering: K-means and Hierarchical methods

Clustering refers to a set of techniques for finding subgroups, or clusters in a dataset. When we cluster observations, we want to divide them into distinct groups so that the observations within each group are very close, whereas the observations in different groups are very different. To do this, we need to define what means to be similar and different for two or more observations. An example of a clustering application arises in marketing. Generally speaking, in marketing campaigns, we have access to a large amount of data about individuals, variables such as median household income, occupation, distance from the closest urban location, and so on. Cluster analysis may be used to conduct market segmentation by finding subgroups

of people that are more sensitive to those types of ads or more likely to buy a certain product. Clustering is a popular technique and therefore exists a great number of clustering methods. In our analysis, we focus on the two best-known clustering approaches: K-means clustering and hierarchical clustering. K-means clustering aims to divide the observations into a predetermined number of clusters. In hierarchical clustering, on the other hand, we don't know how many clusters we want in advance; therefore, we end up with a dendrogram, a tree-like visual representation of the observations that helps us to see the clustering obtained for each possible number of clusters at once.

The K-means clustering method is a technique that divides a data set into K independent, non-overlapping clusters. To use K-means clustering, we have to determine first the number of clusters we require (K); then the K-means algorithm will allocate each observation to one of the K clusters. The idea behind k-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible. One potential disadvantage of k-means clustering is that it requires us to pre-specify the number of clusters.

Figure 3.1 K-means clustering

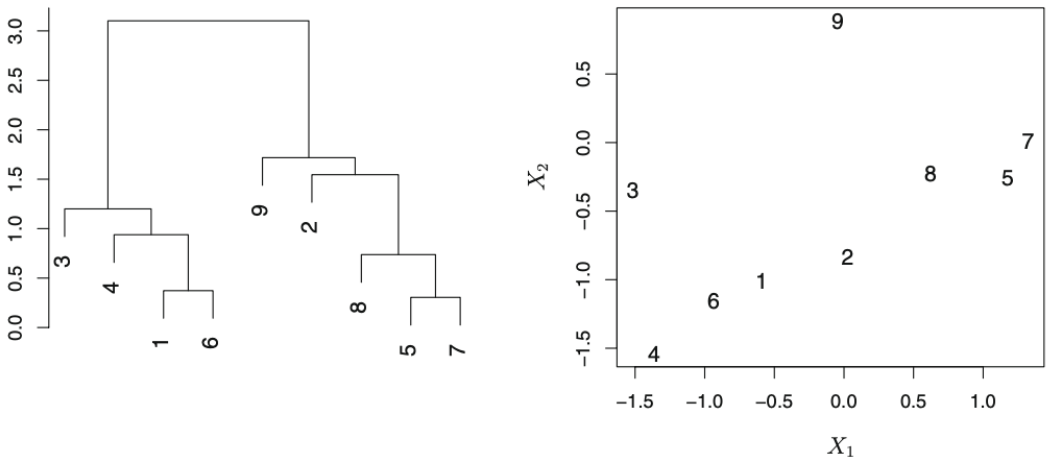


Source: JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., An Introduction to statistical learning, 2013, New York, Springer, page 387.

On the other hand, Hierarchical clustering is a different method that does not force one to pledge to a certain *K* value. Hierarchical clustering has an advantage

over K-means clustering in that it produces a dendrogram, which is a visually appealing representation of the observations. As we see below the dendrogram allows us to assess the distance between observations.

Figure 3.2 Hierarchical clustering



Source: JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., An Introduction to statistical learning, 2013, New York, Springer, page 393.

In figure 3.2 each leaf of the dendrogram represents one observation and as we move up to the tree, some leaves begin to fuse into branches. Lower in the tree fusions occur, the more similar the groups of observations are to each other. On the other hand, observations that fuse near the top of the tree can be quite different.
In figure 3.2 a dataset with nine observations is plotted in a dendrogram representation. On the left side of the figure, we can observe that measurements 5 and 7 are very similar to each other. Instead, observations 9 are no more similar to observations 2 because, even if are near taking into consideration the horizontal distance, they are in two different vertical levels in the high part of the tree. On the right side, we can observe the raw data used to create the dendrogram where is very clear to understand the distance between the observation 9 and 2.

2. Understanding and preparing the dataset

In our analysis, we have tried to understand whether was possible to identify clusters of similar Gimme5 users using unsupervised statistical techniques. First, we have generally studied the dataset in our possession using correlation, after that, we have prepared our dataset selecting the nineteen thousand users previously defined as “Active”. Once defined our dataset, we have detected the proper number of clusters using Elbow and Silhouette analysis, and finally, we have applied PCA and Clustering methods to understand the composition of users within each cluster.

2.1 Understanding the data

In this third chapter, we deep dive into our analysis starting from the datasets provided by AcomeA. In Gimme5, as we have explained in the previous chapter, we can find three different types of users depending on their degree of activity:

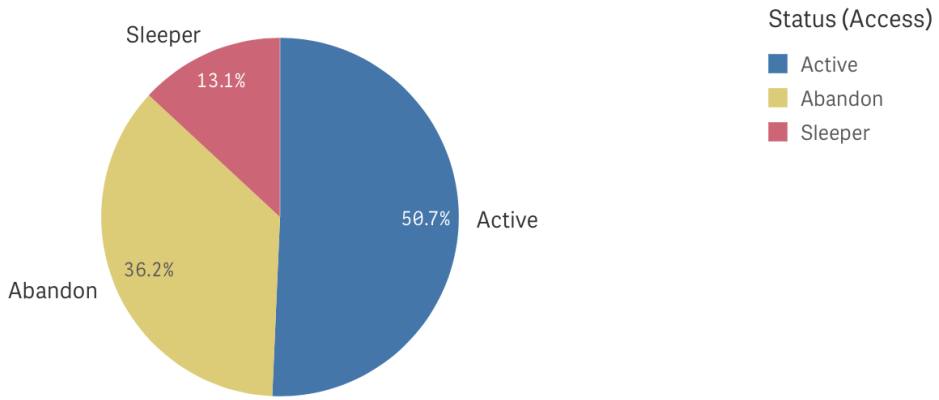
Active users: a user is considered “Active” if his/her last login was within the last 60 days.

Sleeper users: a user is considered “Sleeper” if his/her last login was between 60 and 180 days ago.

Abandon users: a user is considered "Abandoned" if his/her last login was more than 180 days ago.

Moreover, in Gimme5 there is a consistent part of users that have performed only one transaction in the app and after that, they never used the App again. Therefore, we decide to include in our analysis only the users who have performed more than one transaction in the App classifying these users as “Relevant”.

Figure 3.3 Including only Relevant users



Source: Gimme5's dataset

As we can see from the pie charts above, eliminating all the users which have downloaded the App and complete only one transaction, the "Active" users exceed 50% of the total user present in Gimme5.

Once our dataset has been defined, we decided to focus our attention on the "Active" users, because they are the users with more data points to evaluate and the most interesting for Gimme5. Our analysis aimed to discover patterns of behaviors among clusters of Gimme5 "Active" users. In particular, we try to understand if inside Gimme5 exists subgroups of "Active" users that differentiate for demographic variables and saving behaviors.

2.1.1 Variables

In our dataset, we have included eight variables whose six numeric variables, one categoric variable, and one Boolean variable.

Table 3.4 Gimme5's dataset variables

Variable	Type of variable	Description
age	Numeric	It is represented by the age of the users.
gamification_level	Categoric	It is represented by how much a user uses the app to its full potential in a grade from 1 to 5 defined by Gimme5.
average_transaction	Numeric	Average amount transaction for each user calculated in euros.
gender	Boolean	It is represented by the gender of the users where 1 is male and 2 is female.
band_of_wealth	Numeric	It is represented by one of the four bands of wealth where High=1, Medium-High=2, Medium-Low=3, Low=4.
longevity	Numeric	It is represented by the number of days between the first and the last transaction executed on the App.
access_frequecy_yearly	Numeric	It is represented by the fraction between the total number of accesses made by a user and a user's subscription time in years
transaction_frequency_yearly	Numeric	It is represented by the fraction between the total number of transactions made by a user and a user's subscription time in years

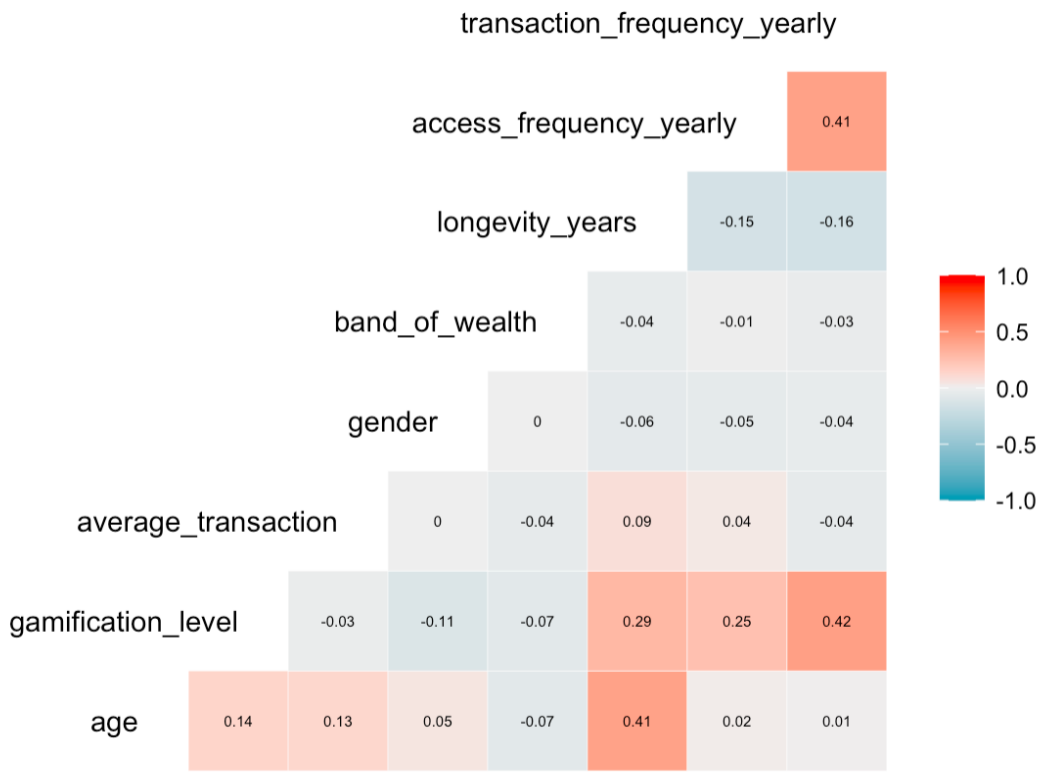
Source: Gimme5's dataset

2.1.2 Correlation

Starting to study our dataset from the correlation we can understand whether there are highly correlated variables. Highly correlated variables can have

implications in unsupervised statistical analysis because those variables may be redundant. When two variables are highly correlated, we may drop one of the two variables and the clustering will not be significantly affected.

Figure 3.5 Correlation in Gimme5's dataset



Source: Gimme5's dataset

A high degree of correlation is defined when the coefficient value lies between +/- 0.50 and +/- 1. As we can see from figure 3.5, in our dataset we don't have highly correlated variables, therefore we can proceed with our analysis without dropping variables.

2.1.3 Preparing the Dataset

Finally, to prepare our dataset for unsupervised learning techniques, we have scaled the data subtracting its mean to each variable and dividing each variable for its standard deviation. Doing this we can have a dataset scaled with the same order

of magnitude for each variable. Now we are ready to apply unsupervised learning techniques.

3. Unsupervised learning techniques in the Gimme5 case

In this final section, the aim was to understand whether exist consistent subgroups of "Active" users in Gimme5. First, we have applied PCA analysis to find a low dimensional representation of the dataset and visually understand whether it is possible to identify clusters of "Active" users. After that, we used Elbow and Silhouette analysis to define the proper number of clusters and performed K-means and hierarchical clustering enhancing and explaining our findings.

3.1 Principal Components Analysis (PCA) in the Gimme5 case

As we explained before, PCA finds a low dimensional representation of a data set that contains as much as possible of the variation. We now explain how these dimensions, or principal components, are found. The first principal component is a normalized linear combination of a set features $X_1, X_2 \ldots, X_i$ that have the largest variance. By normalized we mean that the loadings sum of squares is equal to one because if we set these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance. The loadings are interpreted as the coefficients of the linear combination of the initial variables from which the principal components are constructed. Therefore, given an $n * i$ dataset named *X*, how do we compute the first principal component? Since we are only interested in variance, we assume that each of the variables in X has been centered to have mean zero. We then look for the linear combination of the sample feature values of the form that we see below.

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{i1}X_i$$

Finally, we choose the component with the largest sample variance, subject to the constraint that $\sum_{j=1}^{i} \phi_{j1}^2 = 1$. In other words, the first principal component loading vector solves the following optimization problem.

$$maximize \left\{ \frac{1}{n} \sum_{k=1}^{n} \left(\sum_{j=1}^{i} \phi_{11} X_1 \right)^2 \right\} subject\ to \sum_{j=1}^{i} \phi_{j1}^2 = 1$$

After the first principal component Z_1 of the features has been determined, we can find the second principal component Z_2. The second principal component is the linear combination of $X_1, X_2 \ldots, X_i$ that has maximal variance out of all linear combinations that are uncorrelated with Z_1. Applying PCA to Gimme5's dataset we can see in figure 3.6 that we have eight principal components. This is to be expected because there are in general $min(n-1, i)$ informative principal components in a data set with *n* observations and *i* variables.

To compute the proportion of variance explained by each principal component, we simply divide the variance explained by each principal component by the total variance explained by all seven principal components.

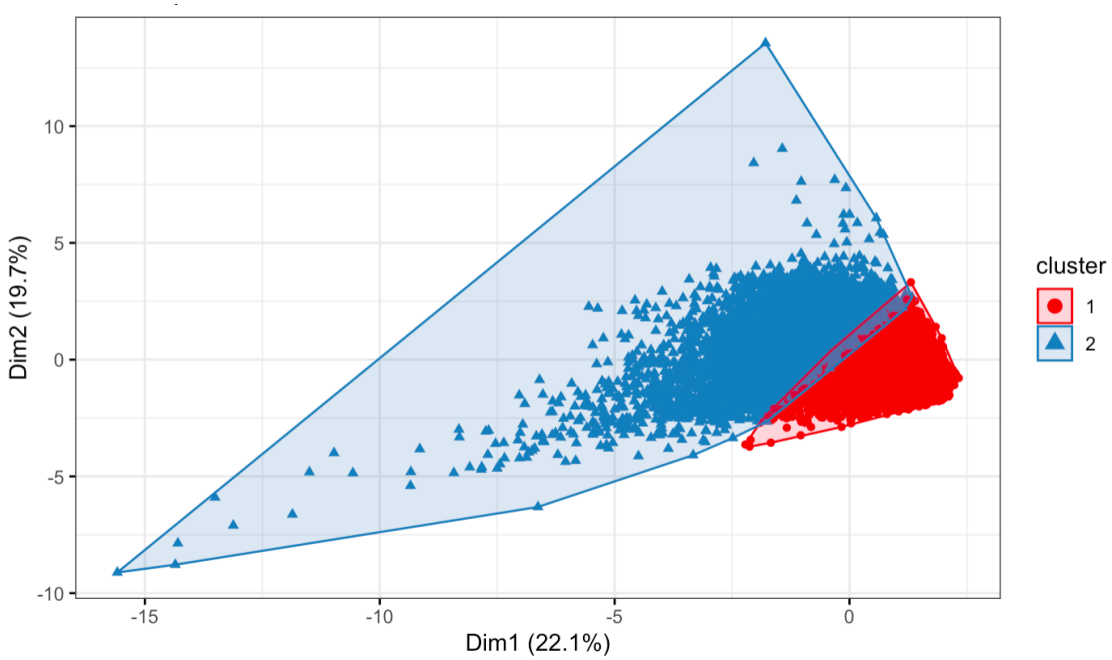
Table 3.6 Principal components dimension

Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8
0.214916	0.193889	0.132884	0.122129	0.121369	0.087809	0.074737	0.052268

Source: Gimme5's dataset

Finally, we are able to use the first two principal components computed to plot our dataset in a Euclidian plan preserving 41.8% of the variation. Observing our analysis, the PCA has reduced the total number of variables (8) to 2 and, as we can see from table 3.7, allow us to individuate two distinct clusters which are not too relevant because the part of variance explained by the PCA is not too large.

Figure 3.7 Principal components dimensions



Source: Gimme5's dataset

As we can see from figure 3.7, two can't be the correct number of clusters because the two clusters aren't clearly identifiable, and the variance explained is only 41.8%. Additionally, the difference is not so marked otherwise it would be possible to see two clusters far apart that do not touch each other and the points on the Euclidean plane would be more centered towards two different centers. This means that we need further analysis to detect the correct number of clusters.

3.2 Clustering: K-means and Hierarchical clustering in the Gimme5 case

Given the PCA analysis, we can visually understand our dataset and hypothesize that the proper number of clusters for Gimme5 is not two. Therefore, to understand whether our assumption is properly formulated, we will use Elbow and Silhouette analysis before applying K-means and Hierarchical Clustering to individuate the correct number of clusters.

3.2.1 Choosing the right number of clusters: Elbow and Silhouette analysis in the Gimme5 case

As we previously mentioned explaining unsupervised statistical techniques from a methodological point of view, usually several clusters, like for example in K-means clustering, are arbitrarily defined. But there are tools like, for example, Elbow and Silhouette analysis that allow us to identify the proper number of clusters.

3.2.2 Elbow analysis on the Gimme5 dataset

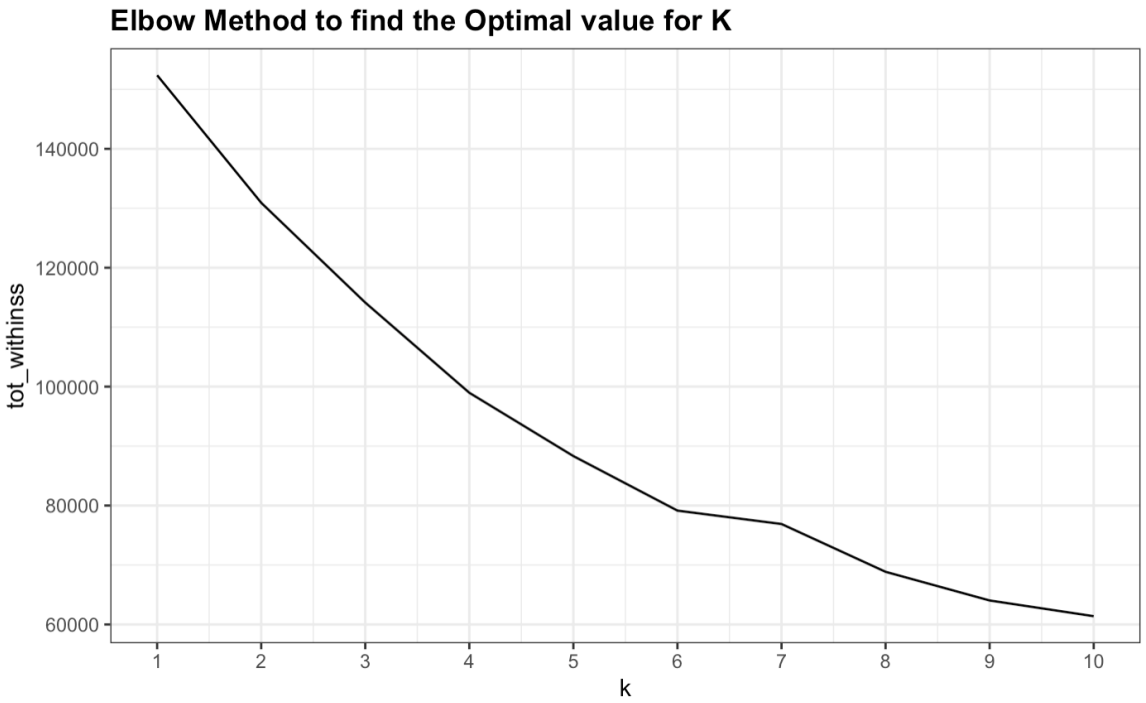
The Elbow method is a way of selecting the optimal number of clusters for K-means clustering. K-means, as we have seen before, is an unsupervised statistical technique that groups data into a specified number (K) of clusters. To do that, we must specify arbitrarily what K to choose, but, at the same time, we need to give a rational to choose the number of K correctly. The Elbow method helps us to give a rational to our decision running k-means clustering on the dataset for a range of values for K (1-10) and then for each value of K computes the sum of squared errors ("SEE"). The distortion score, which we can see on the y-axis in figure 3.8, is computed by the sum of squared error from each point to its assigned center in the Gimme5 dataset.

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x}_i)$$

- x_i data point
- $x\hat{}_i$ assigned center

The idea is that we want a small SSE, but the SSE tends to decrease toward 0 as we increase K (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). Therefore, to detect whether from the Elbow analysis is possible to assess a strong number of clusters, we should be able to see relative maximums (large difference between cluster's SSE) in the line that gradually descends downwards as the number of clusters increases.

Figure 3.8 Elbow method to find the optimal value of K for Gimme5



Source: Gimme5's dataset

In our case are not visible any relative maximum, except for slight variation between cluster 6 and 7, therefore Elbow analysis doesn't give us much help in defining a correct number of clusters, but there are larger used techniques, such as, Silhouette analysis which allow us to better identify the right K.

3.2.3 Silhouette analysis on the Gimme5 dataset

Silhouette analysis is another method to individuate an appropriate number of clusters, in this case using the degree of separation between clusters. The degree of separation between cluster is defined as you can see below:

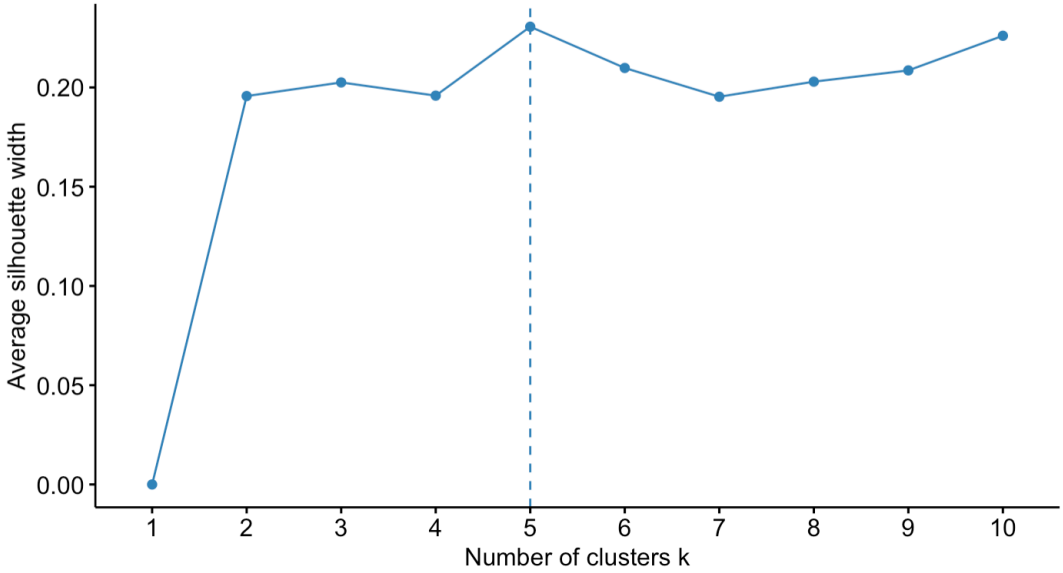
$$s(o) = \frac{b(o) - a(o)}{max\{a(o), b(o)\}}$$

- $s(o)$ is the silhouette coefficient of the data point o;

- $a(o)$ is the average distance between o and all the other data points in clusters to which o belongs;
- $b(o)$ is the minimum average distance from o to all clusters to which o does not belong.

The value of the silhouette coefficient is between -1 and 1. A score of 1 denotes the best meaning that the data point o is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters.

Figure 3.9 Optimal number of clusters with Silhouette method for Gimme5



Source: Gimme5's dataset

Our hypothesis previously mentioned is correct, the appropriate number of clusters is far from two. In the Gimme5 case, the Silhouette analysis shows that the number of clusters with the higher Silhouette coefficient is 5. This means that in our dataset of "Active" users we can individuate five different groups of people with probably different behaviors respect their gender or age.

Another important thing that we can denote from our Silhouette Analysis is that it's true that the highest Silhouette coefficient appears in K=5, but the coefficient is just above 0.20 and this means that could be possible that our clusters will be a

bit overlapped and therefore this makes us hypothesize that the difference between cluster won't be so accentuated.

3.2.4 Applying K-means clustering in the Gimme5 case

To use K-means clustering, first, we have to decide the number of clusters K we would like to have, and then the K-means algorithm will allocate each observation to one of the K clusters identified. Before deep diving into our findings, we explain a little bit more how the K-means analysis is performed from a mathematical point of view. We begin defining $C_1, \ldots, C_k$ as sets containing the observations of each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \ldots \cup C_k = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

2. $C_1 \cap C_{k'} = \emptyset \; for\; all\; k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible. For cluster C_k , a measure WC_k is an amount by which the observations within a cluster differ from each other. Therefore, we want to solve the following problem:

$$minimize_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

In other words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible. But in order to make it actionable, we need to define the within cluster variation. There are many possible ways to define this concept, but by far the most common choice involves squared Euclidean distance. That is defined as

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=i}^{p} (x_{ij} - x_{i'j})^2$$

Where $|C_k|$ defines the number of observations in the kth cluster. In other words, the within-cluster variation for the *kth* cluster is the sum of all of the pairwise squared Euclidean distances between the observations in the kth cluster, divided by the total number of observations in the *kth* cluster. Combining the two equations we give the following optimization problem that defines K-means clustering.

$$minimize_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=i}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

This is a very difficult problem to solve since there are K^n ways to partition n observations into K clusters. Fortunately, an algorithm can be used to provide a local optimum to the K-means optimization problem.

The algorithm executes the following action to find our clusters:

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

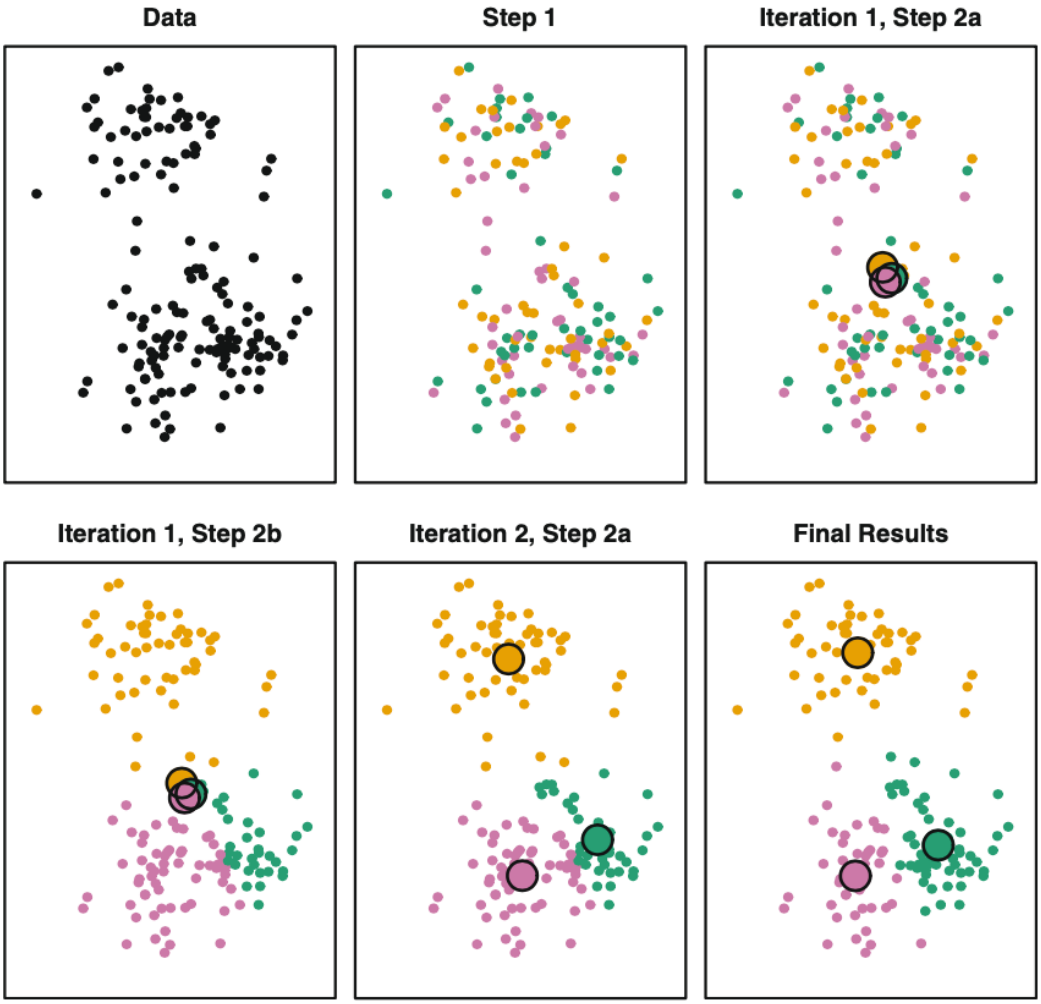
   a) For each of the K clusters, the algorithm computes the cluster centroid. The *kth* cluster centroid is the vector of the p feature means for the observations in the *kth* cluster.
   b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

As we can see from the following identity the algorithm guarantees to decrease the value of the objective at each iteration:

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=i}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=i}^{p}(x_{ij}-\bar{x}_{kj})^2\,,$$

where $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$is the mean for the feature *j* in cluster C_k. Each iteration is composed of two steps, in the first step is calculated the cluster means that for each feature is represented by the constants that minimize the sum-of-squared deviations, and in the second step, the observations are reallocated to the new nearest centroid. This means that the machine learning algorithm will continually learn and improve until the result no longer changes, and a local optimum has been reached.

Figure 3.10 K-means clustering iteration example



Source: JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., An Introduction to statistical learning, 2013, New York, Springer, page 389.

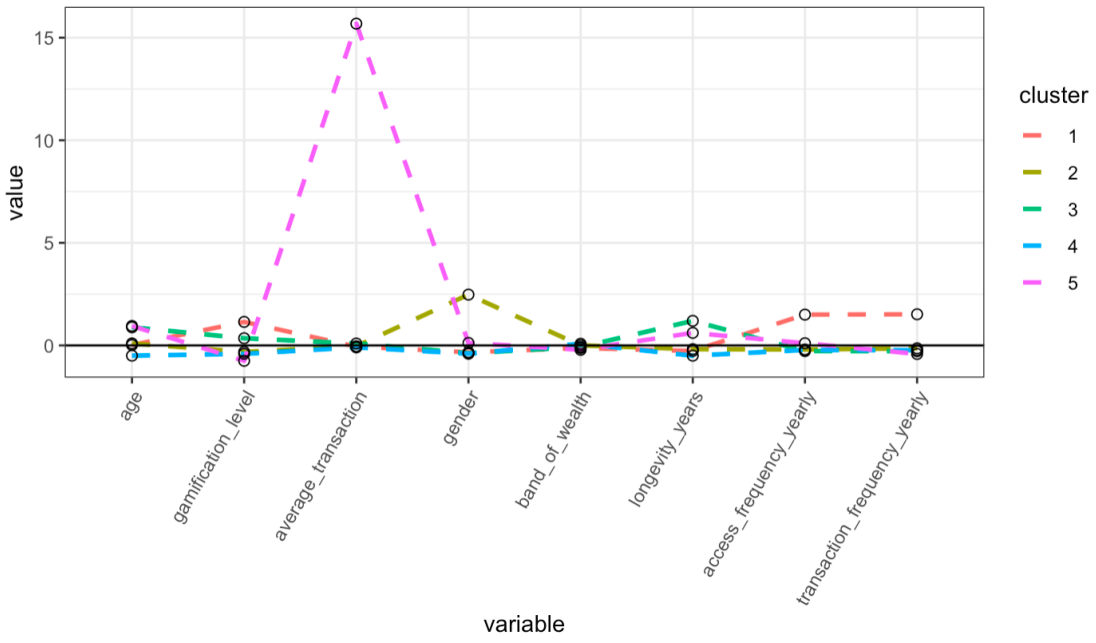
As we can see from figure 3.11, which explains how the algorithm works with an example of K=3, in the first step each observation is randomly assigned to a cluster. In step 2(a), the cluster centroids are computed and are represented by the larger circles. In step 2(b) each observation is assigned to the nearest centroid and once again is performed a new iteration. In the last figure, we can see the results obtained after ten iterations[22].

After this necessary mathematical explication, finally, it's time to test K-means clustering on the Gimme5 dataset.

[22] JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., *An Introduction to statistical learning,* 2013, New York, Springer, pages 385-399.

In our case, we have run our algorithm with a different numbers of clusters from 1 to 5, but K = 5, as our hypothesis and the Silhouette analysis suggested, allows us to show the most proper representation of subgroups in the Gimme5's dataset. Defining K = 5 and set a maximum of 100 iterations our algorithm identifies five clusters containing different number of users. The first cluster contains 2563 users, the second cluster 2560 users, the third 4767 users, the fourth 9114 users, and finally, the fifth only 44 users. Figure 3.11 gives us a first visual overview about the difference among clusters and explains as in the Gimme5 "Active" users there are 5 different clusters of people that, even if are in the same group of the most active users in Gimme5, have five different behaviors and demographic variables. Additionally, by looking figure 3.11 we can notice that all the line except, average transaction amount which we will analyze later, is near to zero for all variables. This makes us assume that the division in 5 clusters is fair, as soon as, the size of each cluster is not too small to make it useless. From a first glance, we understand besides a difference in behaviors and demographic variables among the five clusters, there is a meaningful difference between the average transaction executed in Gimme5 between the fifth cluster and the other four groups.

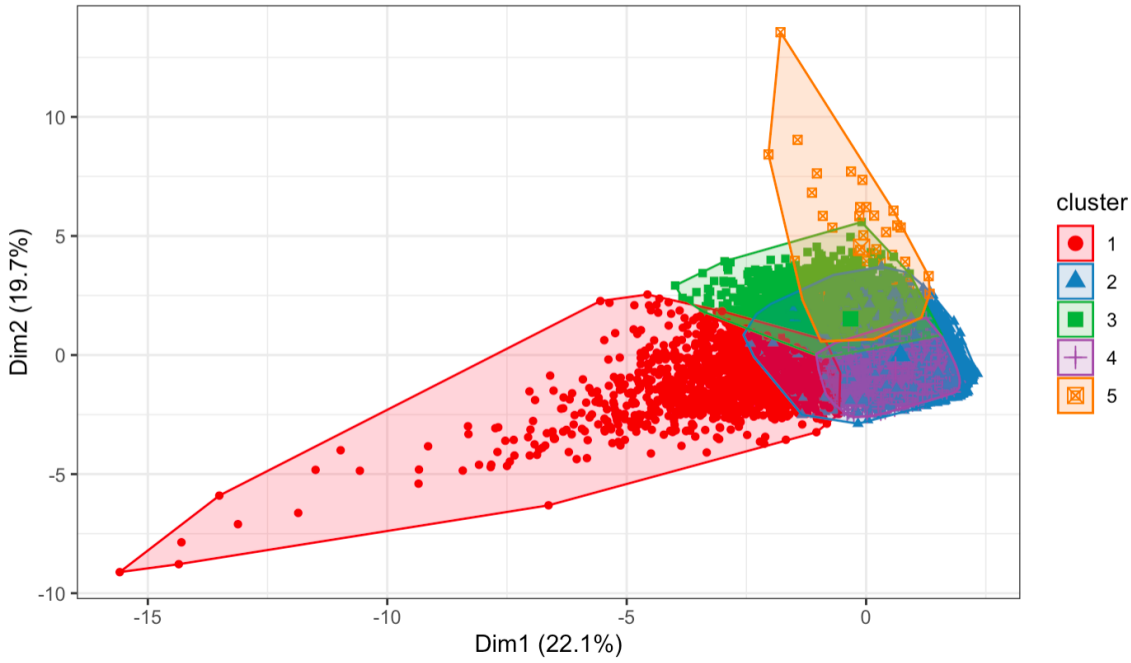
Figure 3.11 K-means clustering with K=5



Source: Gimme5 dataset

In figure 3.11 the data are still scaled, and we are not completely able to understand the differences between clusters, therefore we have decided to summarize all our findings in figure 3.12 and table 3.13. In table 3.13 we have defined our results in a format not scaled highlighting differences between clusters and adding the size of each cluster to underline the magnitude of each group of "Active" users.

Figure 3.12 PCA with K = 5



Source: Gimme5 dataset

Table 3.13 K-means Clustering with K = 5 focus on variables

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
size	2563	2560	4767	9114	44
percentage	13.46%	13.44%	25.03%	47.85%	0.23%
age	36	37	46	30	46
gamification_level	3.5	1.9	2.7	1.8	1.5
average_transaction	54.24 €	58.86 €	97.15 €	43.64 €	4,523.78 €
gender	1.0	2.0	1.0	1.0	1.2
band_of_wealth	1.8	1.9	1.9	2.0	1.7
longevity_years	1.2	1.3	3.3	0.8	2.4
access_frequency_yearly	571	156	138	151	231
transaction_frequency_yearly	122	34	26	29	19

Source: Gimme5 dataset

Now, we can finally comment on our results from a managerial point of view, considering also the magnitude of our analysis which is representative of all the "Active" (almost twenty thousand) users of Gimme5 that represent more than 50% of all the users in Gimme5.

Looking at figures 3.12 and 3.13 we can say that:

- figure 3.12 appears to be overlapping but they are not. The reason why is that the data are multidimensional, but the graph is bidimensional and hence we see them as if they were overlapping but they may be not;
- table 3.13 explains as the fifth cluster is composed only by 44 users but they have an amount of average transaction almost fifty times higher respect the cluster number 3, which is the second for highest amount of average transaction. This made us also think of some kind of error in our calculation, but in the reality checking our dataset with the maximum granularity it is possible to see that in Gimme5 exists a group of "top users" with an average amount transaction very larger in respect all the users present in the App. To understand the economic significance and impact of this small group of users we have decided to multiply the average transaction amount for the yearly transaction frequency to have an approximation of the weight in terms of money transacted in the platform for each cluster. In fact, from figure 3.14 it is possible to see that only forty-four users (0.23% of the total) have contributed with 7.7% of the money invested in the platform by "Active" users;
- the cluster number 1, 2, and 5 are representative of only 27.13% of the Gimme5 users, but together represent 52.3% of capital invested by "Active" users in the platform;
- the bands of wealth are very similar for all the clusters (between 1.7 and 2);

Table 3.14 Money invested in Gimme5 for each cluster with K-means

	total amount	users	weight	Avg capital invested per user
Cluster 1	16,965,766.47 €	2563	34.3%	6,619.50 €
Cluster 2	5,086,262.09 €	2560	10.3%	1,986.82 €
Cluster 3	12,166,231.67 €	4767	24.6%	2,552.18 €
Cluster 4	11,441,710.91 €	9114	23.1%	1,255.40 €
Cluster 5	3,812,502.52 €	44	7.7%	86,647.78 €
Total	49,472,473.65 €	19048	100.0%	2,597.25 €

Source: Gimme5 dataset

- cluster number 1, even if is representative of only 13.46% of the Gimme5 users, is the group with the larger amount of capital invested in the platform by a single cluster (34.3%), the higher gamification level (3.5), the higher access frequency yearly (571 accesses yearly) and the higher number of transactions yearly (122).
- cluster number 2 represents 13.44% in term of users and 10.3% in term of capital invested, but what is interesting is that it is composed only by women;
- cluster number 3 is the group with the second larger amount of capital invested in the platform by a single group (24,6%), the second larger cluster for the amount of user (25,03%), the second cluster for average transaction per user (97.15€), and is represented by the group with the higher longevity (3.3 years) in the platform.
- cluster number 4 is the younger group of users (30 years) and the larger cluster in terms of users (47.85%) but is also the cluster with the lower longevity (only 8 months), and the lower average transaction (43.64%).
- cluster number 5 is the group with the higher average transaction (4523.78€) per user, the less populated cluster (only 44 users), and the group with lower frequency of transaction yearly (only 19);

Summarizing our findings, we can say that there are 3 clusters (1, 2, and 5) that are very important for Gimme5, because even if they represent a minority (only 27.13% of the total users) impact the 52.3% of the capital invested in the platform. Moreover, exist another important minority (cluster 5) to preserve, which is

represented by only 44 users (0.23 % of the total users) but depicts 7.7% of the total capital invested in the platform by active users.

3.2.5 Applying Hierarchical clustering in the Gimme5 case

K-means clustering has the limitation of requiring one to determine the number of clusters K in advance. Hierarchical clustering is a different method that does not force one to pledge to a certain K value. In our second chapter, we have already spoken about Hierarchical clustering from a theoretical point of view. Now, we deep dive into our knowledge about this machine learning algorithm and test it on our Gimme5 dataset.

The term hierarchical refers to the idea that clusters formed by cutting the dendrogram at a certain height must be nested inside clusters formed by cutting the dendrogram at a higher height. Before entering into our analysis, we have to take into consideration an important characteristic of Hierarchical clustering. Let's assume that our findings are representative of a population of people with a 50–50 male-female split, equally distributed among Americans, Japanese, and French. We may envision a situation in which the best division into two groups would divide these citizens by ethnicity, and the best division into three groups would divide them by nationality. The real clusters are not nested in this situation in the sense that the strongest division into three groups would not come from breaking up one of the two groups. Due to such situations, hierarchical clustering can sometimes yield worse or less accurate results than K-means clustering for a given number of clusters.

We begin by defining some sort of dissimilarity measure between each pair of observations. In our case, we will use the classic Euclidean distance and the Ward distance that we will see in the following pages. The algorithm proceeds iteratively, starting out at the bottom of the dendrogram, where each of the n observations is treated as its own cluster. The two clusters that are most similar to each other are then fused so that now there are $n - 1$ clusters. Next, the two clusters that are most similar to each other are fused again, so that now there are $n - 2$ clusters. The algorithm proceeds in this phase until all of the observations belong to one single cluster, and the dendrogram is complete. Finally, in this calculation, we have to take

into consideration how do we determine whether a cluster should be fused with the nearest cluster or not? We have defined a concept of the dissimilarity between pairs of observations, but how do we define the dissimilarity between two clusters if one or both of the clusters contains multiple observations? The concept of dissimilarity between a pair of observations needs to be extended to a pair of groups of observations. This extension is achieved by developing the notion of linkage, which defines the dissimilarity between two groups of observations. The four most common types of linkage are complete, average, single, and centroid.

Complete: Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.

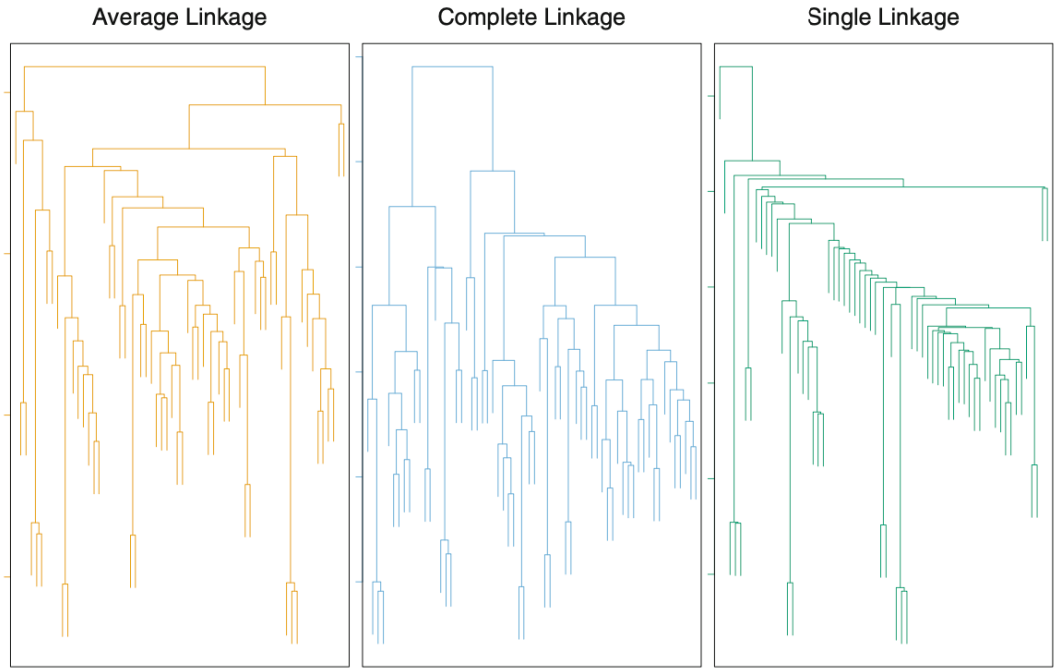
Single: Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.

Average: Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.

Centroid: Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

As we can see from figure 3.15, average (linkage we will use in our analysis), and complete linkage are preferred because produce more balanced dendrograms respect single and complete. On the other hand, a centroid is often used in genomics, but suffers from a major drawback in that an inversion can occur, whereby two clusters are fused at a height below either of the individual clusters in the dendrogram.

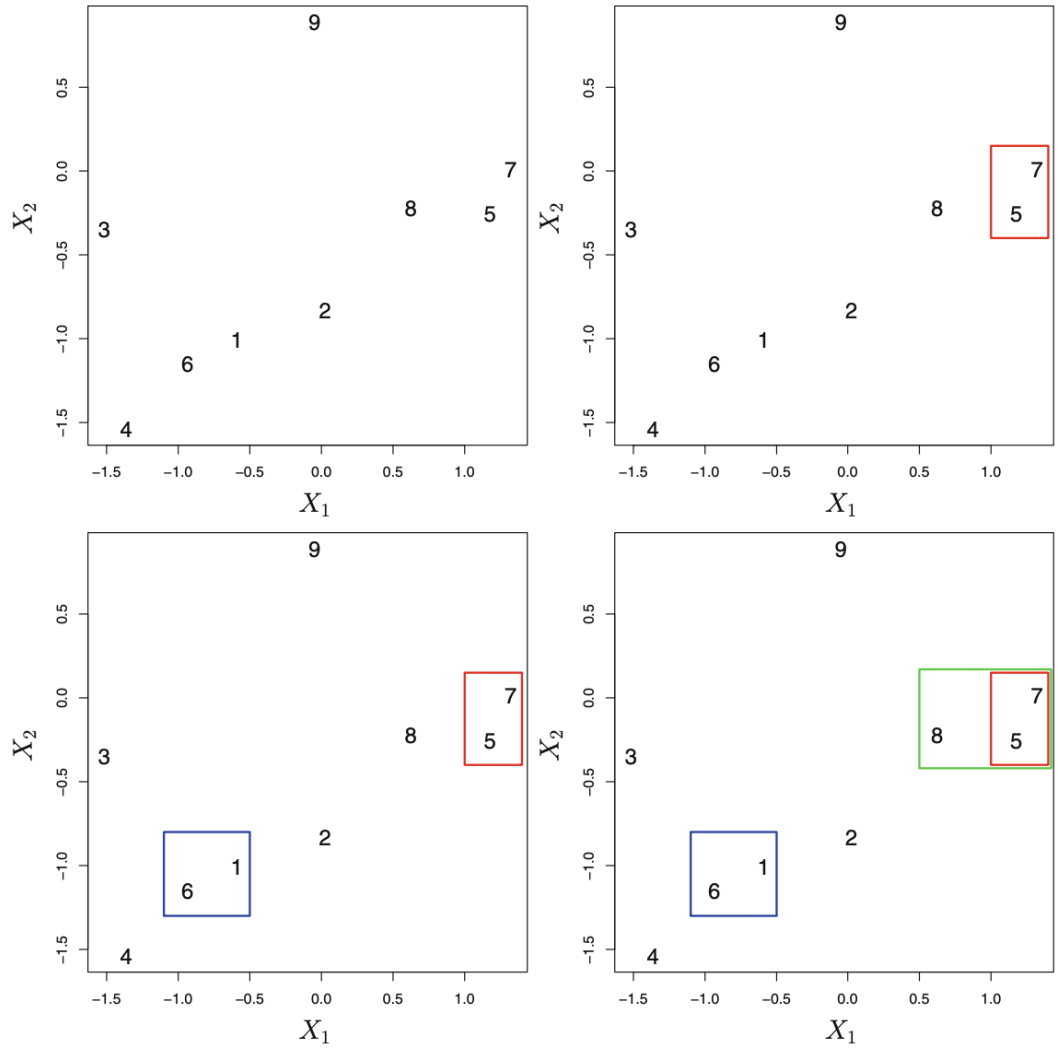
Figure 3.15 Hierarchical clustering different linkage



Source: JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., An Introduction to statistical learning, 2013, New York, Springer, page 397.

Figure 3.16 explains the step in the hierarchical cluster algorithm with complete linkage and Euclidian distance. In the beginning, there are nine distinct clusters, from 1 to 9. In the second step to the top right the two clusters that are closest together, 5 and 7, are fused into a single cluster. After that, in the third step, the two clusters that are closer together, 6 and 1, are fused into a single cluster once again. Finally, in the last graph, we can see the two clusters that are closer together using complete linkage, 8, and the cluster 5, 7, are fused into a single cluster for the last time.

Figure 3.16 Hierarchical clustering iterations



Source: JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., An Introduction to statistical learning, 2013, New York, Springer, page 396.

As we have theoretically explained the Hierarchical Clustering algorithm has different steps, which we can reassume in the following points.

1. Start with n observations and a calculation of all the $\frac{n}{2} = n(n-1)/2$ pairwise dissimilarities (such as Euclidean distance). Consider each pair of observations to be its own cluster.

2. For $i = n, n-1, \dots, 2$:

   a) Examine all pairwise inter-cluster dissimilarities within i to find the least dissimilar pair of clusters (that is, most similar). Combine the two clusters and the difference in height between these two clusters shows where the fusion should be put in the dendrogram.
   b) Compute the new pairwise inter-cluster dissimilarities among the $i = n$ remaining clusters.

After this mathematical explication[23] propaedeutics for our following analysis, it's time to test the hierarchical clustering on the Gimme5 dataset. In our analysis, we will use two different types of distance, the Euclidean distance with average linkage, where we calculate the mean intercluster dissimilarity and pairs those clusters with the smallest average distance between each element in cluster 1 with each element in other clusters, and the Ward distance. The Ward criterion minimizes the total within-cluster variance and, at each step, finds the pair of clusters that leads to a minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centers.
Starting from the Hierarchical clustering with Euclidean distance, the algorithm finds five clusters in the Gimme5 dataset, as the Silhouette analysis suggested us in our precedent analysis, but it's easy to understand looking to figure 3.17 and 3.18, that for the insignificant size of the clusters Euclidean distance it's useless for our analysis.

Table 3.17 Hierarchical clustering Euclidean vs Ward

Type	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Euclidean Distance	19029	12	4	1	2
Ward Distance	4130	4777	2575	5014	2552

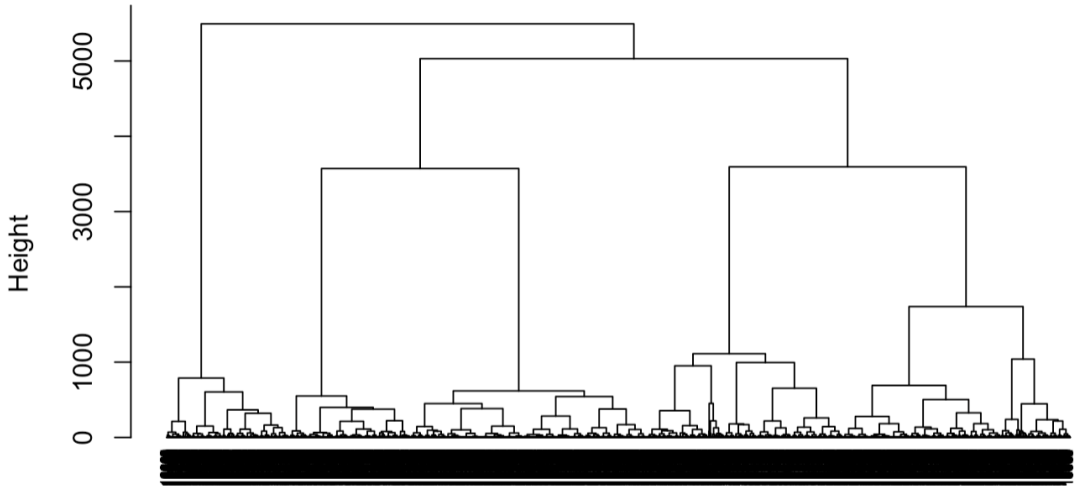
Source: Gimme5 dataset

This means that if we use the average method with Euclidean distance, we have more information loss because some elements in cluster 1 are very different from

[23] JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., *An Introduction to statistical learning,* 2013, New York, Springer, page. 385-399.

the centroid. In other words, there is high within-cluster variance, and we will lose the individual information of these data points. On the other hand, Ward distance gives us five consistent clusters, as easily see from Figures 3.17 and 3.19. This because one of the main reasons why it is common to use the Ward distance method, is that it considers the information loss (Error Sum of Squares, ESS) in deciding which elements and subclusters to the cluster. This means that the configuration that we get through the Ward Distance method, will be the one with the lowest ESS and hence with the lowest individual information loss because within-cluster variance is minimized.

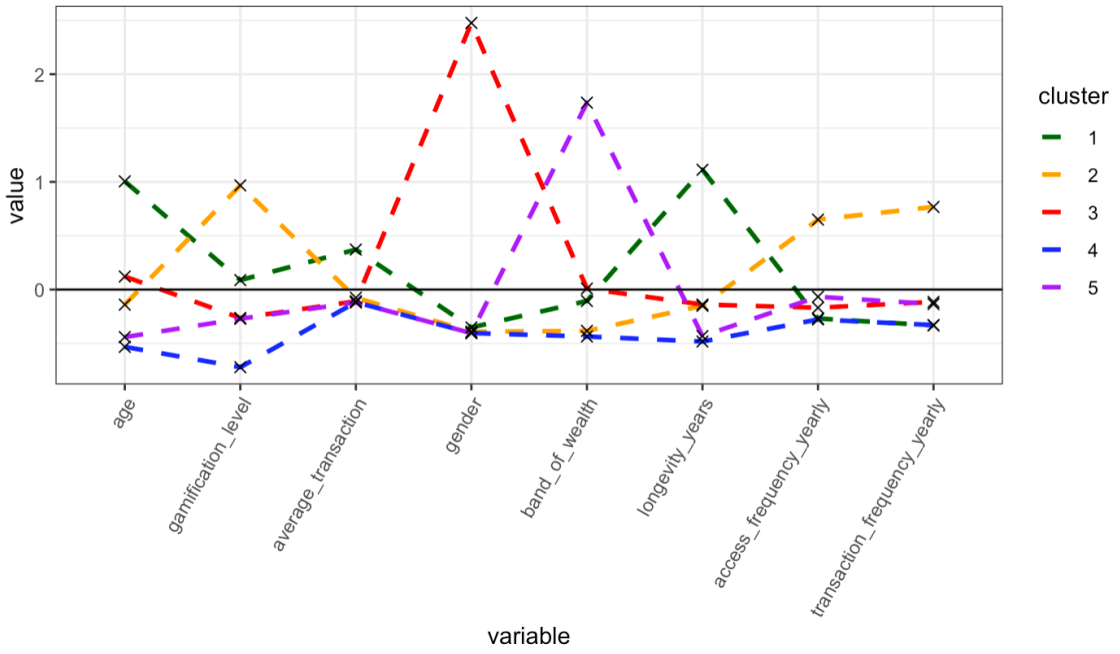
Figure 3.18 Hierarchical clustering with Ward distance



Source: Gimme5 dataset

Therefore, continuing our analysis we focus our attention on the five clusters with Ward distance. Representing our five clusters in a more appealing view (figure 3.20), we can better understand the difference between the five clusters, but we are not able to assess how large these differences are because the data are scaled.

Figure 3.19 Hierarchical clustering with Ward distance



Source: Gimme5 dataset

Therefore, as we have previously done, we have to take back our data to their standard scale to assess how much these differences are large.

Table 3.20 Hierarchical clustering with Ward distance focus on variable

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
size	4130	4777	2575	5014	2552
percentage	21.68%	25.08%	13.52%	26.32%	13.40%
age	40.5	43.7	36.6	32.3	27.0
gamification_level	3.7	1.8	2.0	2.2	2.0
average_transaction	60.8	165.8	39.2	31.9	37.9
gender	1.0	1.3	1.1	1.1	1.1
band_of_wealth	1.7	2.4	1.9	1.8	1.8
longevity_years	1.0	1.1	5.1	1.0	1.0
access_frequency_yearly	134.6	116.7	210.1	623.0	95.7
transaction_frequency_yearly	87.8	57.8	31.4	24.2	13.5

Source: Gimme5 dataset

Now, we can comment on our results from a managerial point of view, taking into account that the Hierarchical clustering has selected the same number of clusters of the Silhouette analysis but, at the same time, it aggregates users in a different way in respect K-means. In fact, it starts from the totality of the users to cluster people in more homogeneous cluster size. This could be interesting to find aggregation in variables such as gender, age, or band of wealth but it could be less

relevant from a business point of view. The reason why is that the Hierarchical clustering algorithm pushes the division towards clusters of similar sizes, and this makes less visible outliers which are useful from a business point of view to detect patterns. For the above-mentioned reason in this part, we will underline only findings that add value in respect to our previously mentioned results.

In fact, looking to figure 3.20 and 3.21 we can say that:

- the five clusters have a more similar size in respect to the k-means because the Hierarchical clustering reason in terms of difference within clusters. On the other hand, the K-means aggregate similar users to find clusters;
- comparing table 3.14 and 3.22 we can see that the total capital invested in the platform is different, but this is normal because clustering techniques return approximation in term of yearly transaction frequency, and the average amount of each transaction;
- Cluster number 2 account for 25.08% of "Active" users and is representative of the 60.1% of the capital invested in the platform, but what is interesting, besides the relevance in term of capital invested, is that this cluster has the lower (medium-high and medium-low) band of wealth (2.4) seen so far.

Table 3.21 Money invested in Gimme5 for each cluster with Hierarchical clustering with Ward distance

	total amount	users	weight	Avg capital invested per user
Cluster 1	22,024,630.21 €	4130	28.9%	5,332.84 €
Cluster 2	45,792,714.92 €	4777	60.1%	9,586.08 €
Cluster 3	3,171,511.85 €	2575	4.2%	1,231.66 €
Cluster 4	3,865,619.07 €	5014	5.1%	770.97 €
Cluster 5	1,308,680.69 €	2552	1.7%	512.81 €
Total	76,163,156.74 €	19048	100.0%	3,998.49 €

Source: Gimme5 dataset

- It's not visible a cluster composed only by women which instead are spread in the other clusters;

- the cluster number 1 are the users with the higher gamification level (3.7), the higher average age (40), and the higher yearly transaction frequency;
- The longevity is very similar in all cluster, except for group number 3 which have longevity almost five times higher respect all the other clusters.

4. Conclusions

In this final chapter, we used both K-means and Hierarchical clustering to segment Gimme5 users: the two methods can be seen as complementary because different algorithms with different logics yield different results. It is up to the researcher to interpret them and gather the right evidence. Specifically, in our case, the K-means clustering has proved to be more useful to find group of users relevant and actionable from a business point of view.

Summarizing our findings, it is sufficient to say that, through the K-means clustering, we have found that three clusters (1, 2, and 5 of table 3.14) that account only for 27.13% of the Gimme5 users, but together represent the 52.3% of capital invested by "Active" users in the platform. In particular, cluster 5 is represented by only 44 users (0.23 % of the total users) but depicts 7.7% of the total capital invested in the platform by active users. On the other hand, Hierarchical clustering divides groups of users with different logics, and we discovered that the cluster with higher age (cluster number 1 of table 3.21) is also the group with the higher gamification level. Moreover, with the hierarchical clustering logics, we have seen that cluster number 2 account for 25.08% of "Active" users and is representative of the 60.1% of the capital invested in the platform, but what is interesting, besides the relevance in term of capital invested, is that this cluster has the lower (medium-high and medium-low) band of wealth (2.4) seen so far.

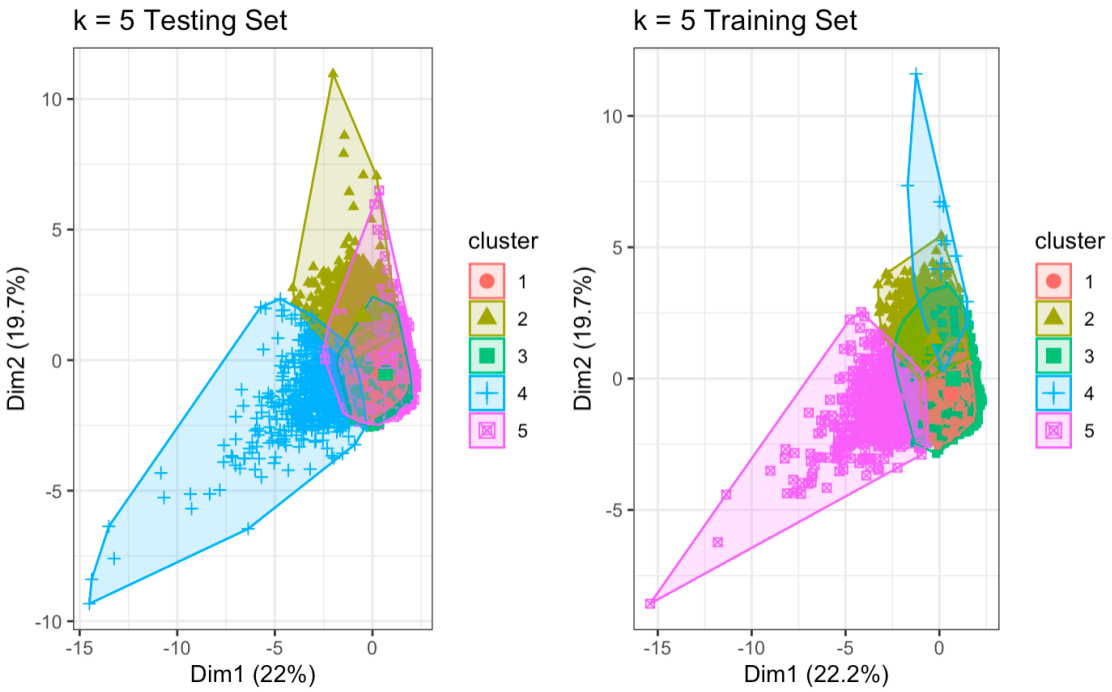
However, for our analysis, the K-means clustering, from a managerial point of view, is more relevant respect the Hierarchical clustering because the cluster found are:

- identifiable: each user's segment can be recognized in the platform;
- sustainable: the size of each cluster is not too small to make it useless and this is especially true for the cluster composed by 44 people (cluster 5 table 3.14) because is a cluster of "top users" that account for the 7.7% of the money invested in the platform;

- accessible: each segment can be reached with dedicated actions, such as special programs for “top users” or specific marketing campaigns;
- actionable: for each cluster can be taken actions to foster towards the Gimme5 business goals.

To give even more consistency to our K-means findings, we have decided to perform a subsample check. The subsample check with PCA consists of dividing our dataset into two parts, training and testing dataset. We performed two different PCA analyses and if the two figures are visually similar, it means that we have individuated the proper number of clusters for the Gimme5 dataset.

Figure 4.1 Subsample check



Source: Gimme5 dataset

As we can see from figure 4.4, the two graphs look similar, therefore we can confirm the consistency of our analysis.

Making our final conclusions, we can also say that Hierarchical clustering, in our case, is less useful from a business point of view but allows us to discover patterns in the population of Gimme5 “Active” users, such as the presence of a cluster (cluster 2 table 3.21) composed with a medium-high/medium-low band of

wealth or a group (cluster 3 table 3.21) with longevity 5 times higher respect the other clusters.

Bibliography

ANDO, MODIGLIANI F., The 'life-cycle hypothesis of saving: aggregate implications and tests, "American Economic Review, 1963, Vol. 53(1), pp. 55–84.

BATSAIKHAN U., DEMERTZIS M., Financial literacy and inclusive growth in the European Union, Bruegel-Policy Contribution, 2018.

DUESENBERRT J., *Income, Saving, and the Theory of Consumer Behavior*, Harvard University Press, United States, 1949.

FRIEDMAN M., *A theory of the Consumption Function,* Princeton University Press, United States, *1957*.

GARGANO A., ROSSI A., There's an App for That: Goal setting and Saving in the FinTech Era, 2020, SSRN Electronic Journal.

GOLLWITZER P.M., Implementation Intentions: Strong Effects of Simple Plans, 1999, American Psychologist, Vol. 54, pp. 493-503.

HARIOKA Y.C., WATANABE W., *Why Do People Save? A Micro-Analysis of Motives for Household Saving in Japan,* The Economic Journal, 1997, Vol 107(442), pp. 537-552.

HERSHFIELD H., SHU S., BENARTZI S., Temporal reframing and savings: A field experiment., 2019, SSRN Electronic Journal.

IVANCEVICH, J., Different Goal Setting Treatments and Their Effects on Performance and Job Satisfaction, Academy of Management Journal, 1977, Vol.20(3), pp. 406–419.

JAMES G., WITTEN D., HASTIE T., TIBISHIRANI R., An Introduction to statistical learning, 2013, Springer, page. 26, 385-399.

KARLAN D., MCCONNELL M., ZINMAN J., Getting to the top of mind: How reminders increase saving. Management Science, 2016, Vol. 62(12), pp. 3393-3411.

KEYNES J.C., *The General Theory of Employment, Interest, and Money,* Palgrave Macmillan, United Kingdom, *1936.*

MODIGLIANI F., BRUMBERG R., Utility analysis and the consumption function: an interpretation of cross-section data, 1954, Post-Keynesian economics. pp 388–436.

PROCHNIAK M., WASIAK K., The impact of the financial system on economic growth in the context of the global crisis: empirical evidence for the EU and OECD countries, 2016, Empirica, Vol. 44, pp295-337.

ROCHER S., STIERLE M., Household saving rates in the EU: Why do they differ so much?, 2015 Discussion Paper.

ROCHER S., STIERLE M., Household saving rates in the EU: Why do they differ so much?, Discussion Paper, 2015, pp 15-20.

RUSSO G., Indagine sul Risparmio e sulle scelte Finanziarie degli Italiani, 2018, Centro di ricerca e documentazione Luigi Einaudi e Intesa San Paolo, Torino.

SOMAN D., ZHAO M., The Fewer the Better: Number of Goals and Savings Behavior, 2011, Journal of Marketing Research, Vol. 48, pp. 944-957.

THALER H.R., Mental Accounting Matters, Journal of Behavioral Decision Making, 1999, Vol.12, pp. 183-206.