Who Are the Fama-French Investors?

An empirical study on Dutch investors' portfolio composition

Supervisor: Dr. Jens Sørlie Kværner

Juan Berasategui Gallego

ANR**:458722** SNR**: 2038913**

Tilburg School of Economics and Management

May 2022

Number of Words: 7011

Abstract

This study researches the association between investors characteristics and their propensity to invest in the long leg of the Fama-French five factor model based on their risk taking and risk bearing capacity, using data from the DNB Household Survey. We show that wealthy, older, university-educated women, who do not own their accommodation, overweight their portfolios towards the long leg. Furthermore, the results are consistent with the portfolio composition, wealth, and age, risk-bearing theories. Lastly, we discuss the implications of the results for the policy makers and the asset management industry.

Key words: Behavioral finance, investor profiling, risk bearing capacity, risk tolerance, Fama-French model, gender, The Netherlands, asset management.

Table of Contents

1. Introduction

On this paper, we build, and make special emphasis, on the research on value and growth investors by Betermier, Calvet, and Sodini (2017). Our results seem to be mostly cohesive with the previous research sustaining their hypothesis that investors with a higher risk-bearing capacity are more prone to tilt their portfolio towards value stocks. Furthermore, the novelties of the study include all the other factors from the 2015 model. In addition, while the Betermier et al. paper focus on risk bearing capacity, we add the element of risk taking. Based on both of these household facets, we hope to offer a more complete picture on who the Fama-French investors are.

The procedure consists of regressing socioeconomical and demographical characteristics on portfolios tilted towards the long leg of each factor of the Fama-French 5 Factor model (FF5FM), therefore we run four different regressions, one per factor, excluding the market risk premium. For the size factor, we found a small association between older investors having a portfolio more tilted towards small caps (β= >0,001). Furthermore, the results indicate that wealthier (β= 0,015), women (β= 0,081), and with university education (β= 0,069) own higher book-to-market stocks. Then, the variables associated with investing in the long leg of the profitability factor consist of gender, women being the investors in the long leg (β= 0,12) again, and whether the investor owns or rents their accommodation, renters being the "Fama-French investors" for this factor (β= 0,17). Lastly, we found an association between wealthier investors and possession of stocks from firms with conservative investing behaviors (β= 0,01).

The values described above stem from portfolios, ranging from 2002 to 2018, retrieved from the DNB Household Survey, an annual questionnaire administered by Centerdata, covering the wealth, income, and identifying data of over a thousand households. Given the messy nature of the database, we convey a rather intensive process of data preparation, including string matching - for which the use of Levenshtein distances was key- editing, cleaning, and tokenization. The programming tools used for the procedure described above are Python, and R.

On the following chapters, we will dive into our motivation and foundations for this study. We will look at the theory in which it is built, and the methods used to convey it. Lastly,

based on the achieved results, we will discuss its implications, and further outlook on the topic.

2. Hypothesis Development

2.1 Fama-French Five Factor Model

The original Fama-French model, built by the University of Chicago academics, had 3 factors. Firstly, Fama and French looked into the traditional capital asset pricing model (CAPM) formula, which predicts expected returns for an asset based on the market premium adjusted to the asset sensibility with the market plus the risk-free rate (Sharpe, 1964), (Lintner, 1965)

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f)$$

$E(R_i) = Expected\ return$
$R_f = Risk\ free\ rate$
$\beta_i = Capital\ asset\ beta\ to\ the\ market$
$E(R_m) = Expected\ return\ on\ the\ market$
$\left(E(R_m) - R_f\right) = Expected\ market\ premium$

Then, Fama and French (1993), based on historical stock returns, came to the conclusion that stocks from small sized firms (small caps) and high book-to-market value firms (value stocks) overperformed the market. Consequently, they decided to add these two factors to the CAPM equation in order to, hopefully, build a better asset pricing model, that is, a model able to explain as many of the returns from a portfolio as possible. The first factor, size, subtracts the big cap historical returns to the small caps historical results, and it is abbreviated as SMB. The second factor, measures the historic returns of value stocks minus those from growth stock, and uses the abbreviation HML (high book-to-market minus low book-to-market). The model that contributed towards Dr. Fama receiving the Nobel Prize in Economic Sciences, can be seen below:

$$R_{(t)} - R_{f(t)} = \alpha + \beta_{(1)}(E\left(R_{m(t)} - R_{f(t)}\right) + \beta_{(2)}SMB_{(t)} + \beta_{(3)}HML_{(t)} + e_{(t)}$$

As seen in "Common risk factors in the returns on stocks and bonds" by Fama & French, 1993.

$t = Time\ period$

$\beta_{(1,2,3)} = Sensibility\ coefficients$

$SMB = Size\ factor$

$HML = Book - to - market\ factor$

$e = Error\ term$

Later in 2015, Fama and French came up with an updated version of their model. This time, the original factors remained, and as in the past they decided to add two more factors, profitability and investment patterns of the firm. The first one refers to the difference between a portfolio composed of stocks with robust profitability and a portfolio of stock with weak profitability, this factor is abbreviated with RMW (robust minus weak). The second represents the difference between portfolio of stocks with a conservative approach towards investing minus a portfolio of stocks investing aggressively, the factor uses the abbreviation CMA (conservative minus aggressive). Below it can be seen the updated Fama-French Model for a portfolio with the two new factors:

$$R_{(t)} - R_{f(t)} = \alpha + \beta_{(1)}(E(R_{m(t)} - R_{f(t)}) + \beta_{(2)}SMB_{(t)} + \beta_{(3)}HML_{(t)} + \beta_{(4)}RMW_{(t)} + \beta_{(5)}CMA_{(t)} + e_{(t)}$$

As seen in "A five-factor asset pricing model" by Fama & French, 2015

$RMW = Proffitability\ factor$

$CMA = Investing\ pattern\ factor$

This Fama and French model (2015) is used to capture the patterns in average stock returns for each one of the factor characteristics. It is able to explain over 70% of the variance in expected returns for size, book-to-market, profitability and investment pattern. Their model is based on the fact that the long leg of Fama-French, that is value, small cap, profitable, and conservative investing stocks outperform the market. Based on this fact, it takes into account the risk for the four factors, besides market risk, and adjust the expected return to it. This model is described as specific to long term investing, since in this time span and with a diversified enough portfolio, investors will be able to ride over short-term volatility.

As we have seen in the descriptions, each factor is built by going long on a portfolio of stocks and shorting another with the opposite characteristic. Therefore, assembling the groups of stocks that Fama and French (2015) go long, results in the Fama-French long leg. Nevertheless, if you group the opposite then one gets the Fama-French short leg.

	SMB	*HML*	*RMW*	*CMA*
Short Leg	*Big-cap*	*Growth*	*Not profitable*	*Aggressive Investing*
Long Leg	*Small-cap*	*Value*	*Profitable*	*Conservative Investing*

For this study, we are interested, and will consequently develop further on the long leg. According to Fama and French (2015), portfolios with these characteristics overperform the market. Therefore, we seek to create a profile based on the characteristics of investors that own stocks in the long leg of the model.

When we discussed the investors characteristics, we will highlight their behavior towards risk-taking. The reason behind this is to be able to generate a hypothesis on who could be the Fama-French investors based on risk appetite. Given the traditional economic wisdom (Brealey & Myers, 1981) we can affirm that risk and returns have a positive correlation, and therefore, given that the long leg from the Fama-French model offers higher returns, it would be fair to think that those who invest in the long leg of the Fama-French five factor model are more risk tolerant.

The Fama-French five factor model, for which the present factors will be used in the study, is not free of critics. Firstly, the model was built using data from the United States, therefore, some academics have shown that specially the factors added in the 2015 model, profitability and investment patterns, do not have an explanatory power, in other significantly different regions, such as South East Asia (Cakici, 2015). Other researchers, corroborate the prior criticism, Foye (2018) tested the five-factor model and noticed that it did better than the three-factor model but just in regions such as Eastern Europe and Latin America, but in practice it did not add much to the Asian market.
Given that, most of the criticism from this last Fama-French model are founded on its explanatory power for Asia, it should not represent a problem for our study.

We investigate a group of independent variables, each variable depicts an individual's characteristic from either of two broader groups: demographics, and socioeconomical conditions. The combination of these two groups has been proved to partly explain financial success (Grable, 2000)

The decision to invest in the long leg of the Fama-French model can be fairly described as a risky decision. Assuming that all investors in the market have full information, we state that to deliberately invest in the long leg, an investor has to be a risk taker, and have a certain risk bearing capacity.

2.2 Variable Description

Definition of Variables

This table summarizes the main variables used in the paper

Variable	Description
age	Age of the respondent
gender	Binary variable, self-reported gender of the respondent (1=man, 0=woman)
log(wealth)	Logarithm of the wealth of the respondent with base 10, wealth is calculated as the sum of all the savings accounts of the respondent
log(income)	Logarithm of the net total income of the respondent with base 10, net total income calculated after taxes from all sources
education	Binary variable, level of completed education of the respondent (1=university education, 0=non university education)
accommodation	Binary variable, whether a respondent owns or not its accommodation (1= owner, 0=not owner)
partner	Binary variable, describes the partner status from the respondent (1=partner, 0=no partner)

2.3 Risk Taking

The originally Greek word, demographics, describes the characteristics of a population. Some examples of these characteristics include, age, race, gender, or partner status. (Salkind, 2010) In our study, the selected demographic characteristics consist of age, gender, education, home ownership, and partner status.

Many academics have looked into the association between age and investing behavior. The usual result is that younger investors have a bigger appetite for risk assets (Brunetti & Torricelli, 2010). In addition, this claim is supported by the research on social science and applied to general risk taking, and not only to financial risk taking. It is found that across the world risky behavior over age follows an inverted U pattern, and that it peaks in the late adolescents (Duell, Steingberg & Icenogle, 2018). Regarding risk bearing capacity,

Regarding the gender characteristic, researchers from different fields seem to come to the same conclusion that men are more inclined to take risks than women (Buunk ,Cobey, Laan, Pollet & Stulp, 2013). Some academics have argued that this difference in based in the disparity in financial education, and that once that both stand on an equally high level of financial education, then there is no difference between genders in financial risk appetite (Hibbert, Lawrence & Prakash, 2013). Other researchers have discussed that men do not take more risks just for the sake of doing so, but because they have a different, likely more overconfident, perception of risk than women (Lee, Miller, Velasquez & Wann, 2013). This difference perception could be explained by the well-researched male overconfidence (Barber & Odean, 2001) (Bengtsson, Persson & Willenhag, 2005).

Another demographic factor that we look into is the education achieved by the respondent. In this study we mark the difference on whether an individual has completed university education or not. Research indicates that education has two effects on financial behavior. Firstly, the more educated seem to have a lower risk aversion (Shaw, 1996), and secondly this same group seems to allocate more capital to the stock market (Black, Devereux, Lundborg & Majlesi, 2018). It is interesting to highlight that the current research on the effect on education on financial risk taking is overwhelmingly focused on financial education, and not

so much in education as a whole. Therefore, the results from this study might be able to throw some light over the topic.

Furthermore, we will look into home ownership, which can arguably be described as one of the most important choices in an individual's life. It is then interesting to ask ourselves, what does making this decision say about a person, or a household. A study written by Rober Shiler in 2007, indicates that during the market boom in that period of time, home ownership was seen by families as an investment opportunity. However, other researchers argument that family formation might be one of the principal reasons behinds the decision of acquiring a new home (Mulder, 2006). The recent boom in the housing market has make it more difficult for most people to enter the market. The relation between this factor and risk aversion, seem to be stronger for older home owners, research shows that wealth increases through house appreciation and therefore the older the household the wealthier it becomes, and the more risk it takes in their financial portfolio (Sodini, Van Nieuwerburgh, Vestman & von Lilienfeld-Toal, 2021).

Lastly, we have the last characteristics of partner status. Which describes whether the respondent lives with a legal partner. The research on this characteristic seem to indicate that marriage has an effect on lowering risk tolerance on individuals (Roussanov, & Savor, 2014). Rui and Sherman (2005), found that financial risk taking among single and married individuals could be ranked from highest to lowest as follows: single males, married males, single females, and married females. Therefore, based on these results, partner status seems to act a mediator variable when paired with gender.

Some studies have shown that savings decisions are heavily influenced by income and educational level of the household (Lindqvist, 1981). Furthermore, some of the drivers boosting saving behavior include age, family size, and length of the planning horizon (Devaney, Anong, & Whirl, 2007). This characteristic, acts slightly different to the ones presented above. Dahlbäck (1991) shows that the relation seems to go on opposite direction, and that savings might be the dependent variable with risk taking as predictor variable. He mentions that the higher the risk aversion, the higher the amount of savings. This research, however, is relatively dated, and we strive to investigate if savings can act as an independent variable to risk taking

Moreover, we want to investigate the income characteristic from each respondent. There seems to be a consensus among academics in that which income is highly correlated with education. Therefore, income would have the same effect on financial risk taking as education, that is, the higher the income, the higher the risk tolerance (Hanna, Gutter, & Fan, 2001).

2.4 Risk Bearing Capacity

Betermier et al (2017) associated risk bearing capacity to age and balance sheet. Then, they proved that as one gets older and one's balance sheet grows (higher risk bearing capacity), one begins to shift their investments from growth to value stocks, that is, from the short leg to the long leg. In another study from Lusardi, Schneider, and Tufano (2010), it was found that when a prospective risk realized, the main coping mechanism used by American households was precautionary savings, followed by extending working hours, and formal and alternative credit. All in all, we model the concept of risk bearing capacity of a household as the liquid resources at disposition in case a financial decision, be it investment, loan, etc. goes south. Our proxy for liquid resources, or wealth, is the capital in saving accounts from the households. Furthermore, in order to build our hypothesis linking risk taking and risk bearing capacity, we regress the rest of the variables chosen for the study as predictors against wealth.

$$Y_{education} = \beta_{age}(age) + \beta_{\log(income)}\log(income) + \beta_{gender}(gender) + \beta_{education}(education) + \beta_{accommodation}(accommodation) + \beta_{partner}(partner)$$

Regression Output

	Coefficients	*Standard Error*	*t Stat*	*P-value*
intercept	2,346	0,161	14,540	0,000
age	0,012	0,002	5,993	0,000
log(income)	0,078	0,020	3,911	0,000
gender	0,172	0,073	2,376	0,018
education	0,241	0,068	3,563	0,000
living	0,179	0,094	1,912	0,050
partner	0,185	0,071	2,594	0,010

As we can appreciate on the regression, there is a clear positive relationship between the rest of the demographic and socioeconomical variables chosen for this study and the wealth of an individual. Moreover, our results are in concordance with those of Betermier et al (2017).

Consequently, our hypothesis for the answer to the question that our research has as title, would lie in an individual who perceives the risk of investing in the long leg of the Fama-French model, and who is willing to take it, basing its decision on its risk appetite and its capacity to withstand the potential negative consequences of the risk. Based on our academic research on both risk taking and risk bearing capacity, we would expect an older, wealthy, university-educated male, with high income, partner, and owning his accommodation as the Fama-French Investor.

3. Methodology

3.1 Levenshtein Distance

A very important part of this study involves natural language processing, in order to turn the households' responses into accurate data. Levenshtein distance is the theory that stands behind the process explained in chapter 3.4.

Levenshtein distance is a theory which measures the needed changes to turn a string into another, these changes can include deletions, insertions, and reversals. The process consists on assigning a "cost" to each one of the operations. Then, after adding all of the needed operations, we will get the Levenshtein distance between the two strings. Furthermore, it is possible to calculate the score between smaller tokenized entities from an original string. For this research, the input and target strings are tokenized into substrings (i.e.: one word = one unique token). Below it can be seen the formula for calculating the recursing Levenshtein distance score between token ***i*** from string ***str₁*** and token ***j*** from string ***str₂*** (Levenshtein, 1966).

$$Levenshtein_{str_1,str_2}(i,j)$$
$$= \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ min\begin{cases} Levenshtein_{str_1,str_2}(i-1,j)+1 \\ Levenshtein_{str_1,str_2}(i,j-1)+1 \\ Levenshtein_{str_1,str_2}(i-1.j-1)+1 \end{cases} & otherwise \end{cases}$$

Furthermore, in order to visualize the quality of the match, it is needed to calculate the Levenshtein distance ratio, which adds the lengths of the input token and best matching token $\boldsymbol{len(str_1) + len(str_2)}$ and then subtracts their Levenshtein distance, and divides over the total length of both strings' tokens.

$$Levenshtein\ Ratio_{str_1,str_2}(i,j) = \frac{[len(i) + len(j) - Levenshtein_{str_1,str_2}(i,j)]}{(len(i) + len(j))}$$

Although this theory dates from the Soviet era, it is still widely used in many academic fields. The use of Levenshtein distance is widely spread, for example in the development of academic plagiarism detection algorithms (Su, Ahn, Eom, Kang, Kim, & Kim, 2008), or in more familiar fields such as contextual spelling correction (Lhoussain, Hicham, & Abdellah, 2015).

The original Levenshtein score, as well as the derived versions, such as the Leveshtein ratio, are remarkably useful since the results can also be interpreted as probabilistic data on which to take decisions for research. Depending on the task and field, one has to choose a threshold which assumes that the matches above are accurate, and the ones below are incorrect. Therefore, given this decision a study conveyed using the same methods might have different results. The decision on where to "make the cut" should then be argumentized taking into account the tradeoff between accuracy and the amount of data.

3.2 Data description

Part of the data used in this research was collected by Centerdata Research Institute. Every year, since 1993, the aforementioned organization surveys over a thousand households, using a questionnaire comprising factors ranging from general household information, through extensive financial data, to health and psychological information among others. The questionnaire, published under the name of DNB Household Survey, has as its main purpose the analysis and understanding of saving behavior present in Dutch households. However, given its broad coverage, the survey has been foreseeably used with many other purposes. Given the focus of this research on profiling the Fama-French Investors in the Netherlands, based on demographic and socioeconomical factors, the following variables were chosen for each year of the survey: Household index , index of the member of the household , year of birth , gender, highest level of education completed , type of accommodation , whether the respondent lives with a partner , total net income, wealth, and portfolio of stocks (*stock1* through *stock10*).

Furthermore, data based on securities information, including identifying, pricing, and fundamental characteristics, was retrieved from S&P Global Market Intelligence's database, Compustat, through Wharton Research Data Services (WRDS), and from Datastream, a Refinitiv service. These data are highly detailed and cover daily information about most financial products. The two security identifying data used were: company name (*company name*) and the International Securities Identification Number (*ISIN*). For the analysis of the securities the following variables were chosen: Market value (*market_value*), lagging market value (*lag1.market_value*), monthly return (*ret*), companies' size (*size*), value premium (*value*), profitability (*prof*), and firms' investing behavior (*inv*).

The timespan for which the empirical study will take place ranges from 2002 until 2018. Said timeline, was selected based on the richness of events between the two dates, including the WorldCom scandal, the 2008 financial crisis, the default of Greece on its debt, but also other positive events to the investors as the continuous dovish monetary policy after 2008, with all time low interest rates as flagship strategy.

3.3 Data Analysis Tools

Once the data is at disposition, it is worthwhile selecting a set of tools needed to edit, clean, format, and lastly, analyze the data.

For the former three required jobs, Python was used. This language designed by Dutch programmer, Guido van Rossum, offers a great variety of libraries that make the process of preparing the data for the analysis simple and intuitive. The packages used in this research include: Pandas, a library aimed at data manipulation and analysis; Openpyxl for writing and reading Excel files; FuzzyWuzzy, for which its name should not sidetrack the readers, from its outstanding string-matching capability; and lastly Levenshtein, which contrary to the previous library, its name does reveal its use for calculation of Levenshtein Distances, and edit operations. All the code was compiled on PyCharm, which is a Czech integrated development environment (IDE) specifically tailored for Python.

For the data analysis, R and Microsoft Excel were the programs of choice. The former commonly is used for statistical computing and, therefore, ideal for the task. The code for this programming language was compiled on its native IDE, RStudio.
The latter, developed by the Washington based company, Microsoft, has a similar goal but offers less flexibility and limited programming capability. It is, however, good enough to carry out the statistical analysis required for the final part of the empirical research.
All the code used for this research is original and written by the author, and Dr. Kværner.

3.4 Data Preparation

As previously mentioned, the preparation process implies data editing, formatting, and selecting among other tasks. The DNB Household survey data, required a larger effort for the above-stated jobs. This is understandable given the quantity of data and range of collection, resulting in some occasions in missing data and change in variable names. It is, however, worth recognizing that for every year Centerdata publishes a codebook indicating all variable changes.

The first and main issue encountered in the survey data, and therefore, the first task for the data preparation process, involved the information on every individual's stock portfolio. Everyone that received the questionnaire, had the chance of imputing the name of up to 10 stocks (*stock1/10*). The problem here is that the collected stock data resulted in mostly incomplete or misspelled security names (e.g.: "univer" instead of "Unilever"). In some occasions, the respondents did use the correct company name ("Unilever"), however, for the used databases this was not enough in order to retrieve the required security data, requiring the ISIN or the registered company name (e.g.: "Unilever PLC"). Our goal consisted on turning the stock imputed by the household component into an official registered stock in order to the be able to then retrieve financial data about it in the Compustat and Datastream databases. Once the inputted stock was already available, the official names with which it would be matched were downloaded from Datastream. In total more than 9000 security names, with their corresponding ISIN, were used as matching targets.

In order to solve the problem mentioned above, we decided to use Python together with the FuzzyWuzzy, and Levenshtein libraries. Firstly, each substring (word) in a string was tokenized for both the input stocks and the matching stocks:

Example

String	*Koninklijke Ahold Delhaize NV*			
Substring	*Koninklijke*	*Ahold*	*Delhaize*	*NV*
Token	1	2	3	4

Then, we used partial token matching based on the Levenshtein distance ratio (Chapter 3.1) between the tokenized input stock name, and the best-matching tokenized matching stock (calculated over all available tokens). In order to get the best matching token, the algorithm iterates over each one of the tokens of the target stocks, at the end it saves the one with the highest score. The script would then return a python list composed of the stock with the best matching token, its ISIN, and the partial token matching ratio, for each one of the input stocks.

Example Output:

Input: *ahold* Return: [“*Koninklijke Ahold Delhaize NV”, “NL0011794037”, 90*]

After carefully reviewing the results of the procedure, it was decided that in order maximize accuracy, even if this meant sacrificing some of the data, only the matches with a ratio of over 88 would be kept for later analysis. Furthermore, the script will then only keep the ISIN.

Once the portfolio of each respondent was clear, it was time to set up the descriptive variables of the members of the households for the later statistical analysis. For this task Python with Pandas was used.

The age variable, being quantitative and in a ratio scale, did not need any further adjustments. However, although the total net income and the wealth variables share the same nature as the previously mentioned variable, it was decided to take the logarithm of both. This decision is based on that such variables tend to be right skewed, were extreme entries (such large incomes, or outstandingly wealthy individuals) distort the results on a regression. When

taking the logarithm of these values what one obtains is more symmetrical and normalized residuals, and as a result, improved validity on the statistical analysis. The rest of the variables relevant to the analysis, are all qualitative, and therefore the values were substituted by dummy, or binary, variables. The assignment of the values was purely based on the most repeated value receiving the 1 (i.e. if there were more men in the dataset, men receive dummy value 1, and women receive dummy value 0).

After this, it was time to classify all of the stocks into the upper or lower tail of each one of the Fama-French factors. We listed all the stocks from the investors' portfolios and then turned the list into a set, so that we would could delete the duplicates. This operation took place using R. Here, if a stock laid on the long tail of one of the factors, it will then receive a 1 for said factor, otherwise, it would receive a 0. Once that all stocks had its rank on the Fama-French model, it was time to move back to Python and build 4 lists, one for each factor, composed uniquely of the stocks which received a 1. Furthermore, a new empty variable under the name of *famafrench_factor* was added to the household data.
The lists were then used to iterate over the household data and build the final files which would then be used for the statistical analysis. This was conveyed by passing 4 equal processes. Each process checked if a respondent portfolio contained, at least, one of the stocks in a particular list. In the cases where the result was positive, the value for the *famafrench_factor* for that household would then be equal to 1, contrarily, if the respondent had only short leg stocks in its portfolio, they would receive a 0 for the newly created variable. Lastly, once the dataframes were completed (*figure 2*), the library Openpyxl, was used to turn what was a comma-separated values file (.csv) into a Microsoft Excel file (.xlsx).

The procedure above marks the end of the data preparation part of this study. The resulting files, one for each of the factors have all the same variables, and they are ready to be analyzed.

3.5 Data Analysis

This section of the study consists of the analysis of the datasets described in the previous subchapter. Here, the hypotheses will be tested, and hopefully after the analysis, it will be possible to answer the research question. This section is the very last step before the cusp of the study, as the results of the analysis work will then build the findings and the further discussion on the topic.

As described in the data analysis tools subchapter, the chosen program for this task was Microsoft Excel, given that it provides a simple and intuitive environment for statistical analysis. Furthermore, using multiple programs, it is positive towards the development of one's program "stack", and consequently, helpful for further research.

Firstly, prior to analyzing the associations between our data and the dependent variable, that is, the tilt in the portfolios towards the long leg of the Fama-French model, we set up a summary statistics table in order to get a superficial understanding of the predictors. The table can be seen below.

Summary Statistics – Independent Variables

Variable	*Mean*	*Median*	*StDev*	*Minimum*	*Maximum*
wealth	44544,87	17260,91	125032,03	0,00	2855000,00
age	57,10	58,00	14,49	18,00	93,00
income	28116,15	27549,51	26576,90	0,00	579583,97
*gender**	0,79	1,00	0,40	0	1
*education**	0,240	0,00	0,43	0	1
*accommodation**	0,893	1,00	0,31	0	1
*partner**	0,783	1,00	0,412	0	1

** Binary variable* *Count = 1963*

In order to answer the question that this research has as title, the selected statistical method was regression analysis. This statistical process available in most statistics-oriented software, is commonly used in many fields, including finance, social sciences, or machine learning, to name a few. Its broad use is based on its capability for identifying and characterizing relationships between a dependent variable and one, or multiple, independent variable (in

some occasions referred as predictors) (Schneider, Hommel & Blettner, 2010). Furthermore, as a results of such method, we are able to draw conclusions based on the coefficients for each one of the predictors in relation to the dependent variable, and on the p-values for both kinds of variables. The former, represents the multipliers from predictors on the dependent variable (Hamilton, Ghert & Simpson, 2015). Moreover, the coefficients are then used as β for each one of the variables in the model equation, depicted in subchapter 2.3. The latter dictates whether the aforementioned association on the response variable, is significant or not, depending on if the resulting P-value lays under the selected α or not, respectively. For this study, an alpha of 0.05 was chosen, and therefore, if the resulting P-value from a predictor is under this level, it will be correct to claim that the coefficient is significant and to reject the null hypothesis (=predictor is not associated with the response variable) to a 95% certainty level.

Assuming the understanding of what a regression and its results implies for an empirical study, the reminding part consist of explaining how the equation was built for this study, and how the results will be interpreted in the following chapter.

Firstly, the dependent variable will be *famafrench_factor*, as a refresher this variable measures whether a respondent is invested in the long leg of a Fama-French factor. As it is desired to understand what characteristics dictate the previous condition, we chose all other variables as predictors. As a result, the regression equation will look as such:

$$Y_{famafrench\_factor} = \alpha + \beta_{age} Age + \beta_{gender} gender + \beta_{education} education + \beta_{accommodation} accommodation + \beta_{partner} Partner + \beta_{\log(\text{income})} log(income) + \beta_{\log(wealth)} log(wealth)$$

nohhold	*nomem*	*stcok1*	*...*	*stock10*	*age*	*gender*	*education*	*accommodation*	*partner*	*log(income)*	*log(wealth)*	*famafrench_factor*
76	1	NL0000009082		NL0011794037	62	1	0	1	0	4.12	2.68	1
1320	2	IT0004513666		NaN	42	0	1	1	1	4.93	3.93	1
1600	1	NL0000009538		NL0006294274	58	1	0	0	1	4.42	4.88	0

On this table, a depiction of how the final data looks can be seen. Note that this is just an example and not necessarily part of the actual data.

Then, when interpreting the results, we will first look at the P-value of the dependent variable, assuming that it will stand below the alpha, then we will go over each of the P-values for the independent variables, if they are considered significant, we will then look at the coefficient, if positive then it would imply that the group with the 1 in such variable (*Figure 1*) are prone to invest in the long leg of the relevant factor for the regression. However, if the β results negative, then this will mean that on the contrary, it is the 0 characteristic that is more susceptible to investing in the long leg.

4. Empirical Results

This chapter consists on showing the results from the regressions and interpreting them. The discussion about them will come then in the last chapter. As mentioned, there are 4 regressions, one for each of the factors of the Fama-French model used in this research. Each subchapter, will showcase the results of one regression.

4.1 SMB Factor

This subchapter carries the title of the factor of size in the Fama-French Model, SMB, which stands for small minus big. Here, the long leg consists of small-cap companies. After running the regression on the respondents that own such stocks in their portfolios, we will try to draw some observations on the drivers of this investing behavior.

Regression Output

$$famafrench_{size} = \alpha + \beta_{\log(wealth)} log(wealth) + \beta_{age}(age) + \beta_{\log(income)} log(income) + \beta_{gender}(gender) + \beta_{education}(education) + \beta_{accommodation} accommodation + \beta_{partner}(partner)$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,155	0,041	3,772	0,000
log(wealth)	-0,004	0,005	-0,735	0,462
age	>0,001	0,001	-2,144	0,032
log(income)	0,001	0,004	0,149	0,881
gender	-0,001	0,019	-0,044	0,964
education	-0,028	0,017	-1,619	0,105
accomodation	0,021	0,024	0,885	0,375
partner	0,029	0,018	1,555	0,119

N= 1963 R^2 = 0.006 4 *Multiple* R^2 =. 0.080.

The variable *age* is the only variable that can be judged significant under a 95% certainty level. The coefficient for said variable is positive, therefore, it is correct to conclude that age does have an association with whether some people will decide to invest in small cap companies. In this case, the results indicates that the older the person, the more prone to do

so. The rest of the variables, are not significant and therefore we cannot reject the null hypothesis.

4.2 HML Factor

This subchapter covers the third factor of the Fama-French five factor model. As a reminder, this factor goes long on high book-to-market, or value, stocks, and goes short the low book-to-market, or growth, stocks.

Regression Output

$$famafrench_{value} = \alpha + \beta_{\log(wealth)} log(wealth) + \beta_{age}(age) + \beta_{\log(\text{income})} log(income) + \beta_{gender}(gender) + \beta_{education}(education) + \beta_{accommodation} accommodation + \beta_{partner}(partner)$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,306	0,062	4,913	0,000
log(wealth)	0,015	0,007	1,965	0,049
age	0,001	0,001	0,381	0,703
log(income)	0,002	0,007	0,240	0,810
gender	-0,081	0,028	-2,856	0,004
education	0,069	0,024	2,830	0,005
accommodation	-0,055	0,033	-1,624	0,104
partner	0,003	0,025	0,133	0,894

N = 1963 R^2= 0.011 *Multiple* R^2= 0.105

In this case, 3 variables can be considered significant, these being: *log(wealth)*, *gender*, and *education.* Under an alpha of 0.05, it is correct to say that the larger the wealthier a respondent, the more its portfolio is tilted towards value stocks. Moreover, women are significantly more likely to invest in growth stocks than men. Lastly, those who have completed university education are more inclined to follow the same path of investing in the long leg of the value factor.

4.3 RMW Factor

This factor, is one of the newly added to the 2015 Fama-French model. It subtracts a portfolio with unprofitable firms to one made out of profitable firms. Looking at the regression results below, it will be possible for the reader to discern which variables might are associated with the investor's decision to buy profitable firms' stocks, and therefore, situate them in the long leg.

Regression Output

$$famafrench_{profitability} = \alpha + \beta_{\log(wealth)} log(wealth) + \beta_{age}(age) + \beta_{\log(income)} log(income) + \beta_{gender}(gender) + \beta_{education}(education) + \beta_{accommodation} accommodation + \beta_{partner}(partner)$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,781	0,063	12,256	0,002
log(wealth)	-0,003	0,007	-0,443	0,657
age	0,001	0,001	-0,132	0,894
log(income)	-0,001	0,006	-0,130	0,896
gender	-0,123	0,028	-4,317	0,001
education	0,051	0,026	1,886	0,059
accommodation	-0,179	0,037	-4,851	0,001
partner	0,027	0,028	0,969	0,332

N = 1963 R^2= 0.022 *Multiple* R^2= 0.148

The observed results for the two variables with a P-value below 0.05 are rather interesting. Similarly, to the previous results, gender seems to play a considerable role in this factor, compared to men, women tend to hold profitable companies' stocks more often. Contrary to size and value, home ownership is significant. Homeowners are less likely to invest in profitable companies compared to those individuals that rent their home. Therefore, these results situate renters and women along the long leg of the Fama-French five factor model.

4.4 CMA Factor

This subchapter covers the last factor of the model, it stands for conservative minus aggressive and describes the investing behaviors of a firm. If a firm does not invest much that makes it a conservative firm and it will lay on the Fama-French long leg, and vice versa.

Regression Output

$$famafrench_{investment\ pattern} = \alpha + \beta_{\log(wealth)} log(wealth) + \beta_{age}(age) + \beta_{\log(income)} log(income) + \beta_{gender}(gender) + \beta_{education}(education) + \beta_{accommodation} accommodation + \beta_{partner}(partner)$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,037	0,029	1,266	0,205
log(weatlh)	0,010	0,003	2,823	0,005
age	-0,001	0,004	-1,376	0,168
log(income)	-0,004	0,003	-1,244	0,214
gender	-0,007	0,013	-0,497	0,619
education	-0,012	0,012	-0,978	0,327
accommodation	0,025	0,017	1,471	0,141
partner	0,011	0,013	0,881	0,378

N = 1963 R^2= 0.008 *Multiple* R^2=0.090

As in the book to value factor, here, we can affirm at a 95% confidence level that the wealthier an individual, the more likely it will be to invest in the long leg of the investment pattern factor of the FF5FM.

4.5 R^2 Value

The R^2 value, or coefficient of determination, is a very important measure in the field of statistics, and even more in regressions. R^2 measures the percentage of the variation on the dependent variable, that can be explained by the predictor variables. Across the regression above, the R^2 was rather low. In some studies, this can be an issue, however, it is not the case for ours. The goal of this paper is to research on whether any of the selected characteristics, has a clear association with Fama-French investors, and then build a profile accordingly. The low R^2 in the study implies that the variability of the predictions is larger, but the direction of the association, which is what is truly important, remains clear. Furthermore, it is important to keep in mind that the driver for this paper are human behavior and portfolio choices, which if anything can be said about them, is that they are remarkably variable.

4.6 Limitation

After the string-matching process, we discussed a trade-off between sample size and accuracy. Given that we decided to go with the accuracy path, the limitation of this study resides in the sample size. This reflects on the results above, there is the possibility that with a larger sample size we could have found more of the variables in the regressions to be significant, although this is not a given. However, if we expanded the data by accepting a lower Levenshtein ratio, we could have attributed wrong stocks to the investors. A potential solution for the limitation in our research, could be to maintain our small sample, but to use a different statistical approach such as Bayesian methods, that is, assuming that we could prove a good prior probability distribution, in our case, probability distribution of *variable* investing in the long leg of the Fama-French factor.

5. Conclusion & Recommendation

5.1 Conclusion

Now that we have the results for the empirical study, it is time to look back on our hypothesis from chapter 2, and evaluate if they correspond to our findings.

The hypothesis consisted on our expectations for the Fama-French investor profile. Based on the available literature and driven by the relation of the risk bearing capacity theory developed by Betermier et al (2017) and the risk-taking theories described in conventional economics, we expected the profile of Fama-French investor to be an older, wealthy, university-educated male, with high income, partner, and owning his accommodation. The results, did support this hypothesis on its majority. One of our main findings is that we reaffirm the results from Betermier et al, proving that wealthier investors are more prone to have value companies with conservative investing patterns in their portfolios. In addition, contrary to our expectations, it was the women who mostly overweighted their portfolio towards the long leg of the Fama-French model. Moreover, compared to owners, those who did not own their accommodation seemed to tilt their investments to the long leg of the model.

Furthermore, assuming that the Brealey and Myers (1981) findings on risk and return can be applied to this setting, we would expect that given that the long leg of the Fama-French model has higher returns and overperforms the market, it would, in the same fashion, imply higher risks. Derived from the empirical study results, it is difficult to say whether the long leg does, in fact, imply higher perceived risk, however, given the results the resulting Fama-French investor -a wealthy, older, university-educated women, who do not own their accommodation- can be judged as having a high risk bearing capacity.

If we uniquely focused on the Fama-French five factor model as the financial performance measure, it would then be fair to claim older university-educated wealthier women, who do not own their accommodation, as the best investors in The Netherlands, based on the results for this study, and the academic literature that describes the long leg of the 2015 model.

5.2 Recommendation

We can affirm that our results are relevant to The Netherlands, however, it would be interesting to convey further research on other countries, especially in those with a non-western culture. Consequently, it would be possible to compare the results and see if our profile can be generalized to the rest of the world. If this was not the case, then this could open the door to further work on public policy on the other countries to try to uprise financial performance among women.

Moreover, as mentioned during the hypothesis development, we believe that the selected variables form a rather complete picture of each one of the respondents. However, we do not close the door to carry out research on the same topic adding new variables and characteristics. For example, it would have been interesting to include, if it existed, a variable explicitly describing the risk tolerance of each individual in order to be able certainly declare a risk perception for the long leg of the Fama-French model.

Our last recommendation is directed towards the private sector, more specifically to asset management firms, and particular investors. Despite the fact that effort is being made on closing the gender gap in the asset management industry, women are still underrepresented as a fraction of public fund managers, representing only the 11% (Morningstar, 2021). Consequently, given our results of women overperforming the market, having equal representation on asset management would not only benefit women, but also the investors that they are representing.

References

Alladi, A., & Dibenedetto, G. (2021, March 16). *The percentage of U.S. female fund managers is exactly where it was in 2000*. Morningstar, Inc.

American Psychological Association. (n.d.). *Socioeconomic status*. American Psychological Association. Retrieved April 27, 2022, from https://www.apa.org/topics/socioeconomic-status

Bengtsson, C., Persson, M., & Willenhag, P. (2005). Gender and overconfidence. *Economics Letters*, *86*(2), 199–203.

Betermier, S., Calvet, L. E., & Sodini, P. (2017). Who are the value and growth investors?. *The Journal of Finance*, *72*(1), 5-46.

Brad M. Barber, Terrance Odean, Boys will be Boys: Gender, Overconfidence, and Common Stock Investment, *The Quarterly Journal of Economics*, Volume 116, Issue 1, February 2001, Pages 261–292,

Brealey, R. A. & Myers, S. (1981). *Principles of Corporate Finance*. McGraw-Hill.

Cakici, N,(2015). The Five-Factor Fama-French Model: International Evidence. 1-51

Cobey, K. D., Laan, F., Stulp, G., Buunk, A. P., & Pollet, T. V. (2013). Sex Differences in Risk Taking Behavior among Dutch Cyclists. Evolutionary Psychology.

Dahlbäck, O. (1991). Saving and risk taking. *Journal of Economic Psychology, 12*(3), 479–500.

Devaney, S., Anong, S., & Whirl, S. (2007). Household savings motives. *Journal of Consumer Affairs*, *41*(1), 174–186.

Duell, N., Steinberg, L., Icenogle, G. *et al.* Age Patterns in Risk Taking Across the World. *J Youth Adolescence* **47,** 1052–1072 (2018).

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22

Foye, J. (2018). A comprehensive test of the Fama-French five-factor model in emerging markets. *Emerging Markets Review*, *37*, 199–222.

Grable, J.E. Financial Risk Tolerance and Additional Factors That Affect Risk Taking in Everyday Money Matters. *Journal of Business and Psychology* **14,** 625–630 (2000).

Grefenstette, G., & Tapanainen, P. (1994). What is a word, what is a sentence?: problems of Tokenisation.

Hamilton, D. F., Ghert, M., & Simpson, A. H. (2015). Interpreting regression models in clinical outcome studies. *Bone & joint research*, *4*(9), 152–153.

Hanna, S. D., Gutter, M. S., & Fan, J. X. (2001). A measure of risk tolerance based on economic theory. *Journal of Financial Counseling and Planning*, *12*(2), 53.

Hibbert, A. M., Lawrence, E. R., & Prakash, A. J. (2013). Does knowledge of finance mitigate the gender difference in financial risk-aversion? *Global Finance Journal*, *24*(2), 140–152.

Lee, Kevin and Miller, Scott and Velasquez, Nicole and Wann, Christi, The Effect of Investor Bias and Gender on Portfolio Performance and Risk (2013). The International Journal of Business and Finance Research, v. 7 (1) pp. 1-16, 2013

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710)

Lhoussain, A. S., Hicham, G., & Abdellah, Y.(2015). Adaptating the levenshtein distance to contextual spelling correction. *International Journal of Computer Science and Applications*, *12*(1), 127-133.

Lindqvist, A. (1981). A note on determinants of household saving behavior. *Journal of Economic Psychology*, *1*(1), 39–57

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance*, *20*(4), 587.

Lusardi, A., Schneider, D., & Tufano, P. (2010). Households@ Risk: A Cross-Country Study of Household Financial Risk. In *American Economic Association Meeting*.

Marianna Brunetti & Costanza Torricelli (2010) Population age structure and household portfolio choices in Italy,The European Journal of Finance, 16:6, 481-502,

Mulder, C.H. Home-ownership and family formation. *J Housing Built Environ* **21,** 281–298 (2006).

Nikolai Roussanov, Pavel Savor (2014) Marriage and Managers' Attitudes to Risk. Management Science 60(10):2496-2508

Salkind, N. J. (2010). *Encyclopedia of research design* (Vols. 1-0). Thousand Oaks, CA: SAGE Publications, Inc.

Sandra E Black, Paul J Devereux, Petter Lundborg, Kaveh Majlesi, Learning to Take Risks? The Effect of Education on Risk-Taking in Financial Markets, *Review of Finance*, Volume 22, Issue 3, May 2018, Pages 951–975

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits. *Journal of Cross-Cultural Psychology*, *38*(2), 173–212.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt international*, *107*(44), 776–782.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425.

Shaw, K. L. (1996). An empirical analysis of risk aversion and income growth. *Journal of Labor Economics*, *14*(4), 626–653.

Sherman, H., & Rui, Y. (2005). The effect of gender and marital status on financial risk tolerance. *Journal of Personal Finance*, *4*(1), 66–85.

Shiller, R. (2007). Understanding recent trends in house prices and home ownership. *National Bureau of Economic Research,* Pages 89-123

Singh, Prabhjot & Dhawan, Sumit & Agarwal, Shubham & Thakur, Dr. Narina. (2015). Implementation of an efficient Fuzzy Logic based Information Retrieval System. EAI *Endorsed Transactions on Scalable Information Systems*. 2. 10

Sodini, P., Van Nieuwerburgh, S., Vestman, R., & von Lilienfeld-Toal, U. (2021). Identifying the Benefits from Homeownership: A Swedish Experiment. *National Bureau of Economic Research*, Pages 89-123

Su, Z., Ahn, B. R., Eom, K. Y., Kang, M. K., Kim, J. P., & Kim, M. K. (2008, June). Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm. In *2008 3rd International Conference on Innovative Computing Information and Control* (pp. 569-569). IEEE.