INTERNATIONAL JOURNAL ACADEMIC RESEARCH IN BUSINESS AND SOCIAL SCIENCES

INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

HR MARS MANAGEMENT ACADEMIC RESEARCH SOCIETY
EXPLORING INTELLECTUAL CAPITAL
www.hrmars.com
ISSN: 2222-6990

Investigation of Individual Investment Preferences with K-Mode Cluster Analysis Based on Socio-Demographic Characteristics

Ayse Yildiz, Emine Ebru Aksoy

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v10-i7/7415 DOI:10.6007/IJARBSS/v10-i7/7415

Received: *21 April 2020,* ***Revised:*** *26* May *2020,* ***Accepted:*** *14* June *2020*

Published Online: 28 July 2020

In-Text Citation: (Yildiz & Aksoy, 2020)

To Cite this Article: Yildiz, A., & Aksoy, E. E. (2020). Investigation of Individual Investment Preferences with K-Mode Cluster Analysis Based on Socio-Demographic Characteristics. *International Journal of Academic Research in Business and Social Sciences*, *10*(7), 280–295.

Vol. 10, No. 7, 2020, Pg. 280 - 295

http://hrmars.com/index.php/pages/detail/IJARBSS	**JOURNAL HOMEPAGE**

INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

HR MARS MANAGEMENT ACADEMIC RESEARCH SOCIETY
EXPLORING INTELLECTUAL CAPITAL
www.hrmars.com
ISSN: 2222-6990

Investigation of Individual Investment Preferences with K-Mode Cluster Analysis Based on Socio-Demographic Characteristics

Ayse Yildiz, Emine Ebru Aksoy
Assoc. Prof. Ankara Haci Bayram Veli University, Faculty of Economics and Administrative Sciences,
Department of Business Management, Ankara, Turkey
Email: ay.yildiz@hbv.edu.tr, emine.aksoy@hbv.edu.tr

Abstract: In recent years, data mining methods have been frequently used in financial investment decisions and the most common one is clustering. The aim of this study is to demonstrate the feasibility and availability of clustering method to propose for the most suitable investment alternative according to the socio-demographic characteristics of individual investors. In order to achieve this goal, the questionnaire method was conducted with 332 individual investors regarding their socio-demographic characteristics with investment preferences which were specified as gold, interest and stock options. Since all variables are categorical, the k-mode cluster algorithm, which is the extension of the k-mean algorithm, was implemented. The results of the analysis indicated that, apart from the risky stock alternative, only two investment options are suitable for these investors and they are risk-avoiders. Another result revealed that only gender and marital status are factors affecting investment preferences. This result will be beneficial for investment advisors to make investment suggestions to individual investors by focusing on these factors. These findings prove that clustering method can be applied effectively in determining suitable investments for individuals, since similar results have been obtained with previous studies.

Keywords: Investment Preference, K-Mode Cluster Analysis, Socio-Demographic Factors, Cluster Analysis, Risk Attitude.

Introduction

The economic units (individuals, enterprises and states) whose incomes are more than their expenses have the opportunity to save money as a result of the surplus income. The economic units, which evaluate their savings in various investment instruments, want to be compensated for giving up their current consumption. How the savings will be evaluated by the owners and their impact on economies have a great importance. Since keeping the savings under the pillow will not contribute to the national economy, it is wanted to attract surplus funds to the financial markets and contribute to the country economy. In this case, it is necessary to determine the factors that affect investment preferences and take necessary measures accordingly.

Individual investors can assess the savings themselves or get help from experts. In this context, professional investors, who manage the savings of small investors on the basis of high returns and low risk, are called institutional investors. While making an investment decision, the preferences of the investors are important for both the investors and institutional investors. Especially in attracting savings to financial markets and to bring them into the using of economy, there are a variety of factors that affect individual investors' preferences, and some of these factors affect the investor's decision directly or indirectly. Factors affecting individuals' investment decisions may be environmental factors or individual factors. Factors such as the situation and development of national economies, the development of institutions-rules-tools in the financial markets and the environment in which individual investors live can be considered as environmental factors. Individual investors with a heterogeneous structure have differences in socio-cultural, economic structures, risk perspectives and financial literacy situations from each other, so their investment decision unlike each other. In addition, individuals are affected by behavioral, emotional and psychological factors while making investment decisions (Wameru, Munyoki, & Uliana, 2008). The process of making investment decisions of individuals and determining the factors that affect these decisions are an important necessity for making predictions for the future, for the development of the national economy, and financial markets. On the other hand, the estimation of investment instruments to be selected by individuals according to their characteristics or investment advisers' investment instruments proposals, it is of particular importance to know the factors that affect the investment decision and the ways in which these factors influence.

Individuals decide on the basis of the risk and return of the investment instruments while making an investment decision in general. In traditional financial theories, it is stated that investors are generally rational and they aim to achieve the highest return against the lowest risk. However, it can be observed that investors do not always display rational investor behavior. When making an investment decision, individual investors are affected by behavioral tendencies, make systematic errors and can show non-rational behaviors (Aydın & Ağan, 2016). The fact that investors move away from rationality may cause fluctuations in stock markets (Faroog, Afzal, Sohail, & Sajid, 2015, s. 63) Therefore, determining the factors affecting the investor decision in terms of the stability of the markets has a great importance (Kapoor & Prosad, 2017). On the other hand, the determination of these factors is inevitable in terms of the validity of the plans of the governments and financial planners and the stability of the decisions to be made. Accordingly, the examination of both individual and behavioral factors that influence the investment choice of the investors is becoming an important necessity in order to explain the investor behavior.

The fact that the investors are guided by the investment advisors or investors give their own investment decisions correctly depends on the investors conditions. For this reason, an investment alternative that is attractive for an investor may not be considered suitable for another investor. In this respect, grouping of similar investors according to their profiles and making investment advises for the groups may facilitate investment advisers' work as well as guiding investors to make their own decisions. In this way, the funds of the savings get rid of staying under the pillow and can be channeled to the economy and contributed to the financial markets. The main purpose of this study is to analyze individual investors by clustering analysis and to separate them into homogenous groups according to their characteristics. As a result of clustering, it is aimed to be a guide in directing investments based on the individual characteristics of the group. The using of clustering analysis to determine investment preferences based on the socio-demographic characteristics of individual investors makes this study different from previous studies. Within the scope of this purpose, in the next section, the literature reviewed to reveal the factors affecting the investment decision of individual investors. The third section introduces the clustering method focusing on the k-mode algorithm explaining with the example. Fourth section includes the data set with methodology and the fifth section indicates the findings and evaluations of these findings. With the conclusion and recommendations part the study is completed.

Literature Review

There are many factors affecting the investment decisions of individual investors. The determination of these factors is of great importance both for directing savings to investment and for making investment alternative selection decisions. There are several studies on this subject from past to present. In these studies, many different factors that are expected to affect individual investors were analyzed. In these analyzes, factors such as socio-economic and demographic characteristics, financial literacy, attitudes and behavioral approaches of individual investors have been discussed.

The majority of the studies have conducted in order to determine the factors affecting the investment decision of individual investors have been related to stock investments. In the most of these studies, the effects of the socio-economic and demographic characteristics of the investors on the stock investment decision were examined. In this context, in the study conducted by Usul, Bekci, and Eroglu, (2002) found that young people, high-income people, men and high-educated people were looking more willing and longer-term to in choosing stock and risky investment instruments. In a similar study, Goetzmann and Kumar (2008) concluded that young, low-income, and low-educated individual investors were inadequate in portfolio diversification in the US. On the other hand, a study by Akbulut and Kaderli (2009), in which the results were interpreted from a different perspective, stated that investment in the stock market seems to be a gamble as the level of education decreases. Moreover, they stated that the quality of the investment was taken into consideration as the level of education increased and the portfolio sizes increased with the increase of the age and income of the investors. In addition, Yesildag and Ozen (2015) and Geetha and Vimala (2014) studies showed that age, occupation, education and income factors affected investment decisions. The other study that takes these studies one step further, Gumus, Koc,and Agalarova (2013) showed that ambitious behaviors and habits can be effective in investment decisions with these factors. Angi, Bekci and Karatas (2016) found that males, married couples, and individual with higher education level preferred the stock investments more. On the other hand, Dogan, Yildiz and Topal (2016), who examined the relationship between investors' personality, demographic characteristics, investment decisions and risk perceptions, showed that investment preferences changed according to demographic characteristics. According to their results, women took into account the return, the men took into account the risk and diversification, the risk-taking tendency increased as the level of education decreases, and the singles preferred more risky investment instruments.

Investors who are financial literacy and have the knowledge and skills to evaluate the data of the enterprises to be invested also make investment decisions according to various indicators related to the firms. In the study conducted by Ali and Rehman (2013) in order to examine the factors affecting stock decisions making of individual investors in Pakistan, they concluded that firms' dividends, price trends and volatility, position in the market, the source of the recommendation, reputation, social performance, and media visibility had a significant impact. On the other hand, a study was conducted by Obuyami (2013) to examine together socio-economic factors, demographic factors and the data of the firms to be invested. According to the results obtained in this study, the most important factors affecting the investment decision were the past prices of stocks, expected share division, capital increase, dividend policy as well as age, gender, marital status and education level. In a similar study was done by Shafi (2014) and it was stated that gender, age, income and education factors as well as company-specific dividends, ownership structure, accounting data, expected returns were also effective factors. In a different study, Lodhi (2014) examined the effects of financial literacy, accounting information, openness to experience and information asymmetry in the decision of investment. It was shown that financial literacy and accounting knowledge reduced the information asymmetry of the investor and facilitated investment in risky investment tools, that risk aversion increased with the experience and age, and dividends were preferred rather than capital gains. On the other hand, a study that demonstrated that the investment decision was affected by the level of income, past investment experiences, experts and other investors' opinions and financial stability was made by Islamoglu, Apan and Ayvali (2015).

One of the most important factors affecting the investment behavior of individuals is the attitudes and behaviors of individuals. In this context, Demir, Akcakanat and Songur (2011) discussed individual investor behavior in terms of behavioral finance and revealed that investors made systematic errors and even if they knew the rational solution, they did not apply this method. Furthermore, they stated that environmental factors such as media and friends influence on the choices of investors, and that these processes that turn into herd behaviors may cause anomalies in the markets. Collu (2018) determined that the advice of the brokerage firm, the willingness to take risks for high returns, the fast becoming rich, the advice of friends and family, the opinions of the majority of the shareholders and the need for diversification the factors were important. In terms of investment risk, Sunden and Surette (1998) found that investors who were married choose the risk-free investment. Faroog, Afzal, Sohail and Sajid (2015) conducted a study to demonstrate the effects of risk aversion and corporate governance application levels on the investment decision, and they determined that the increase in the corporate governance level had positive and the risk aversion had a negative significant effect in making the investment decision. (Aksoy, 2018) examined the factors that affect investment risk preferences and found that young-men-single-childless people, and people with high income or wealth can take the risk of investment more easily. In this study, individual investors are divided into homogeneous groups applying cluster analysis according to their socio-demographic characteristics and it is aimed to make the most appropriate investment alternatives for them depending on these characteristics.

Unlike prior studies used classical methods such as regression analysis, the objectives of this study illustrates how clustering analysis can be applied to answer the following questions:

- Which socio-demographic characteristics with risk taking ability of the investors influence their investment preference.
- What are the most preferable investment alternatives by investors based on their socio-demographic characteristics.

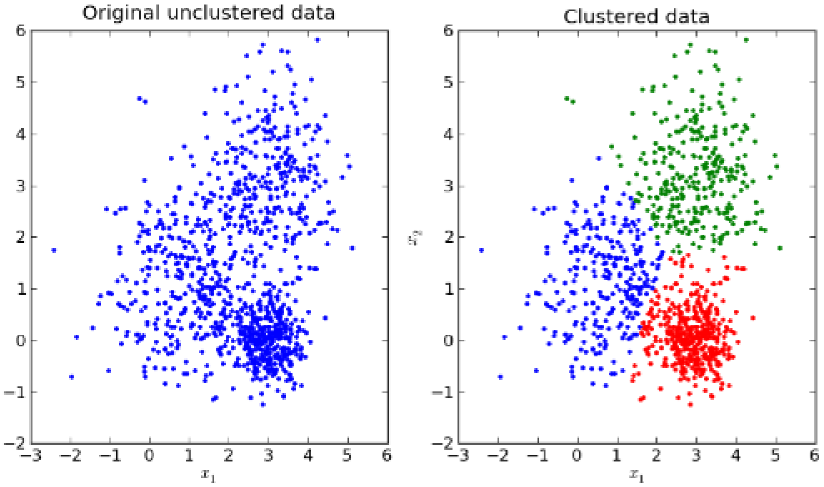
The Method

In this section, clustering analysis is explained briefly with variable selection and clustering method applied.

Clustering Analysis (Subjective Segmentation)

A cluster is a collection of data objects which are similar (or related) to one another within the same group (i.e., cluster) and rather dissimilar (or unrelated) to the objects in other groups (i.e., clusters). Cluster analysis (or clustering, data segmentation) is a given a set of data points, partition them into a set of groups (i.e., clusters) which are as similar as possible. Thus, cluster analysis helps the analysist partition massive data into groups based on its features as seen in the Figure 1.

Figure 1. Clustering the Data



Source: (Patlola, 2020)

The cluster analysis is an unsupervised learning. In an unsupervised analysis there is no dependent variable, predefined classes or objective function. In other words, the relationships between observations are not known in advance. There is only data set and the aim is to get some pattern in this data set and use this pattern to make better decision.

There are mainly two applications for the usage of the cluster analysis:

- The first one is as a stand-alone tool. The aim of this usage is to get insight about data distribution or pattern. The application might be similar products or customer segmentation. Also, it can be useful analysis to detect dynamic trend for clustering stream data and trend or patterns.
- The second application is as a pre-processing tool. This time it is an intermediate step for other analysis or algorithm. Based on intermediate step there are different objectives for application of the clustering analysis. For example, generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing etc. (Han, Kamber, & Pei, 2012).

Variables Selection

At the beginning of the clustering process the decision maker or analyst must select the appropriate variables for clustering. Indeed, selecting the appropriate variables is one of the most fundamental steps in clustering process. One or two irrelevant or improper variables may distort a useful cluster solution. Variable selection is typically based on implicit assumptions and judgements and experience, as well as intuition from the experts (Tuma, Decker, & Scholz, 2011).

Clustering Method Selection

Clustering methods mainly can be classified into hierarchical and non-hierarchical methods (a partitioning method).

Hierarchical Methods

Hierarchical methods do not identify a set of clusters directly. Rather, they identify relationships among objects on the basis of some measure of similarity. Hierarchical solutions are preferred if a wide range of alternative clustering solutions is to be examined and the sample size is moderate (under 300-400).

The method can be classified as being either agglomerative (bottom up) or divisive (top down), based on how the hierarchical decomposition is formed. The hierarchical clustering technique can be visualized using a dendrogram. A dendrogram is a tree-like diagram that records the sequences of mergers or splits as illustrated in Figure 2.

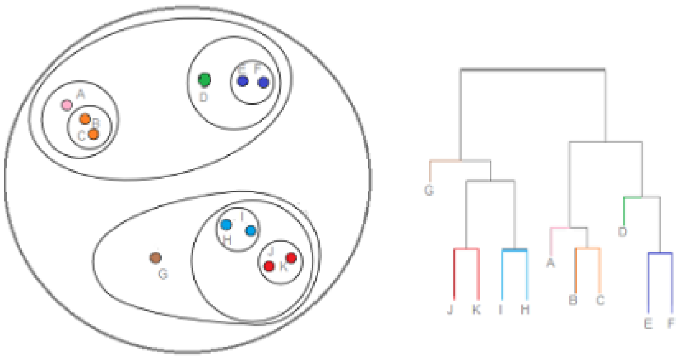


Figure 2. Hierarchical Clustering with Dendrogram
Source: (Hierarchical Clustering / Dendrogram: Simple Definition, Examples, 2020)

However, the application of hierarchical clustering methods is not always justified in the context of market segmentation. First, the use of hierarchical methods presupposes an underlying hierarchy among the objects or respondents to be clustered, which is usually not available in customer segmentation. Second its application becomes difficult with large sample size due to the large number of distance computations. These drawbacks can be handled by applying non-hierarchical clustering methods.

Non-hierarchical method (A partitioning method) - K Modes Algorithm

Non-hierarchical methods derive a partitioning of the sample into clusters directly from the raw data. In partitioning method first creates an initial set of k partitions, where parameter k is the number of partitions to construct. It uses an iterative reallocation technique that attempts to improve the partitioning by moving objects from one group to another. They are preferred when the true number of clusters is known and initial seed points can be specified according to some practical, objective or theoretical basis (Tuma, Decker, & Scholz, 2011). This algorithm processing the data is illustrated in Figure 3.

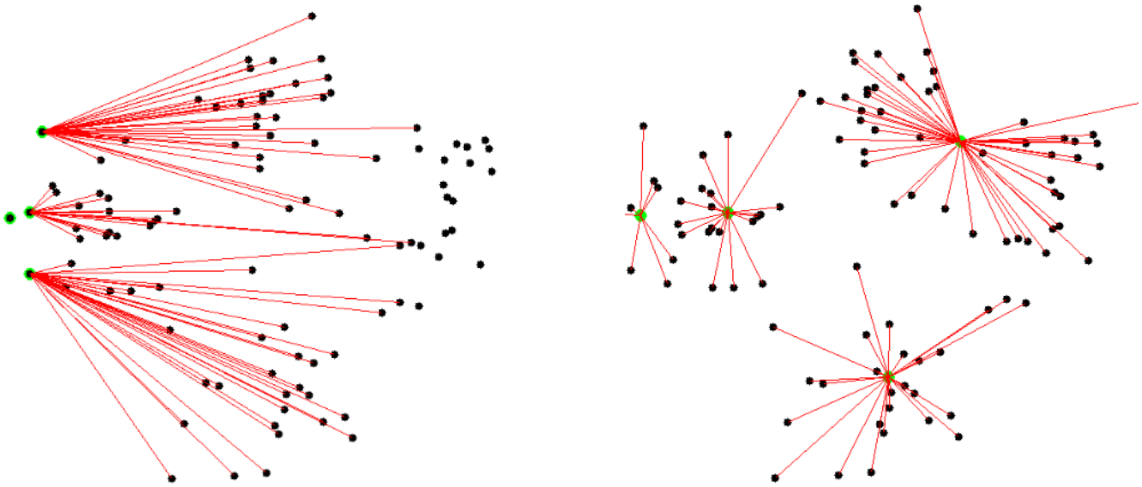


Figure 3. K-Means Algorithm Process
Source: (Chauhan, 2020)

The question is how can we measure distance measurements and based on these distance measurements how can we cluster the data? Distance measurements are different based on the data types. For quantitative data sets, there many different measurements such as Euclidean distance, Manhattan distance etc. and k-means algorithm can be proceeded for this dataset. However, k-means clustering algorithm fails to handle datasets with categorical attributes because of difficulty of calculating means and distances. To handle for this problem, mode measurement is used for mean and for distance problem dissimilarity measurements are developed. For this approach, firstly it is required to convert multiple category attributes into binary attributes (using 0 and 1 to represent either a category absent or present) and treats the binary attributes as numeric to cluster categorical data and then conduct k-means algorithm.

K-modes algorithm approach modifies the standard k-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure. This algorithm uses modes and updating modes with the most frequent categorical values in each of iteration of the clustering process. These modifications guarantee that clustering process converges to a local minimal result. The steps for k-mode algorithm can be defined as follows (Huang Z. , 1998; Khan & Amir, 2013):

The first step is to measure the distance applying a simple matching dissimilarity for categorical objects. To calculate these measures between two objects X and Y described by m categorical attributes, the distance function in -k-modes is defined as the following equations (Huang J. Z., 2009).

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \qquad (1)$$

Where

$$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j. \end{cases} \quad (2)$$

Here, xj and yj are the values of attribute j in X and Y. This function is often referred to as simple matching measure or Hemming distance. The larger the number of mismatches of categorical values between X and Y is, the more dissimilar the two objects.

In this way it is guaranteed that one can obtain only two possible values of similarity: one for the matched categories, and the second for unmatched categories. The frequencies of the contingency table can be used to understand of these dissimilarity measures. Let the symbols from table 1 be given as a,b,c and d to illustrate the frequency of the matched and unmatched categories (Rezankova, Praze, & Vysoka, 2009).

Table 1. The frequency of matched and unmatched categories

Cat. of object xj	Categories of object yj	
	1	0
1	a	b
0	c	d

For dissimilarity measures based on the distance matrix there are three coefficient calculation which are Simple Matching, Jaccard and Dice coefficient. The simple matching is the common and easy one. Therefore, to give some idea about k-mode algorithm simple matching coefficient calculation which is shown in equation 3 is applied.

$$\text{SM}(x,y)= \frac{b+c}{a+b+c+d} \quad \text{(Simple Matching)} \quad (3)$$

The second step is to determine the cluster centers. In k-modes clustering, the cluster centers are represented by the vectors of modes of categorical attributes. If data set has m categorical attributes, the mode vector Z consists of m categorical values (z_1, z_2, …., z_m), each being the mode of an attribute. The mode vector of a cluster minimizes the sum of the distances between each object in the cluster and the cluster center. The third step is update clustering process conducting frequency-based method (Huang J. Z., 2009, s. 246-247).

Dataset and Analysis Results

With this perspective, the clustering algorithm is followed as defined above. The first step is to determine appropriate decision variables. The decision variables (attributes) are selected as socio-demographic characteristics with investing preferences which are interest, stock and gold investments. These socio-demographic characteristics include gender, marital status, education, job (occupation), age, and income. In the study, the data set of the investor characteristics are obtained via online data collection method, and 332 individual investors are reached. Due to some problems occurred while answering the questions, 312 individual investor answers are used for analyzing. Before clustering these investors based on their socio-economic characteristics, it is appropriate step to investigate the dataset by obtaining descriptive data analysis results.

Descriptive Data Analysis Results

To get some idea for the dataset obtained with questionnaire, descriptive data analysis results are illustrated in Table 2.

Table 2. Descriptive Characteristics (Attributes) of Investors

Characteristics (Attributes)	**Categories**	**Frequency**	**Frequency %**
Gender	Female	139	0, 45
	Male	173	0,55
Marital status	Single	161	0,52
	Married	151	0,48
Education	High _School	73	0,23
	University degree	*190*	*0,61*
	Graduate School	49	0,16
Job (Occupation)	Government Labor	35	0,11
	Private Sector	75	0,24
	Government Officer	69	0,22
	Businessman	16	0,05
	Other	117	0,37
Age	Very Young (18-29)	107	0,34
	Young (30-39)	121	0,39
	Old (40-49)	62	0,20
	Very old 50-59	22	0,07
Income (TL)	Very Poor (1000-2500)	46	0,15
	The Poor (2501-3500)	69	0,22
	Medium Income (3501-4	86	0,28
	High Income (4501-5500	57	0,18
	The Rich (5501-6500)	31	0,10
	Very Rich (6501 and over	23	0,07
Investment Alternatives	Interest	167	0,54
	Stock	12	0,04
	Gold	133	0,42

As seen from the Table 2, gender and marital numbers are not very different from each other. On the other hand, for education attribute the proportion of university degree with the rate of 61% is more than other options. Regarding job attribute, businessman is less and the other alternative is more dominant. For the age attribute, young investors have more frequency with %39 and very old investors have less frequency with 0.07%. While the proportion of very rich investors have the lowest percentage with 0.07%, medium income investors constitute the major part of the investors with 28

%. The results regarding investment alternatives indicates that interest is the most preferable alternative with 0.54 % and gold is close to the interest alternative with 0.42% while stock is not preferable option for Turkish investors. To see clearly the investment alternatives based on the socio-demographic attributes, the results can be reconstructed as illustrated in Table 3.

Table 3. Descriptive Attributes of Investors Based on the Investment Alternatives

		Investment Alternatives			**Total**
Characteristics	Categories	Interes	Stock	Gold	
Gender	Female	74	2	10	139
	Male	93	10	70	173
Marital status	Single	97	5	59	161
	Married	70	7	74	151
Education	High _School	40	3	30	73
	University	102	6	82	190
	Graduate School	25	3	21	49
Occupation	Government Labor	19	4	12	35
	Private Sector	44	3	28	75
	Government Officer	38	2	29	69
	Businessman	2	1	13	16
	Other	64	2	51	117
Age	Very Young (18-29)	51	5	51	107
	Young (30-39)	73	3	45	121
	Old (40-49)	29	3	30	62
	Very old (50-59)	14	1	7	22
Income	Very Poor (1000-2500)	21	3	22	46
	The Poor (2501-3500)	30	3	36	69
	Medium Income (3501-450	53	2	31	86
	High Income (4501-5500)	32	0	25	57
	The Rich (5501-6500)	17	2	12	31
	Very Rich (6501 and over)	14	2	7	23

The first point in the Table 3 is that the stock column values are very low compared to columns showing other alternative options. This point shows that people do not prefer stock investment regardless of their socio-demographic characteristics. When the numbers are examined more closely, it can be readily seen that the interest option with gold investment is dominant compared to stock option. Hence, it can be concluded that in general. Turkish investors do not find the stock option attractive which is risky. In other words, Turkish investors can be defined as risk aversion investors.

The other remarkable outcomes acquired from Table 3 are the following. The first outcome is that while the majority of women prefer interest there is no apparent difference in men's preferences between interest and gold. The second outcome is that single people prefer interest more, while married people prefer both interest and gold investments. The third outcome is that investors prefer interest alternative except businessman whose selection is gold. The fourth outcome is that young and very old investors prefer interest more than gold while very young and old investors prefer interest and gold equally. The last outcome is that the preference of the medium income, rich and very rich investors is interest rather than gold while no significant difference between interest and gold is for the other income groups.

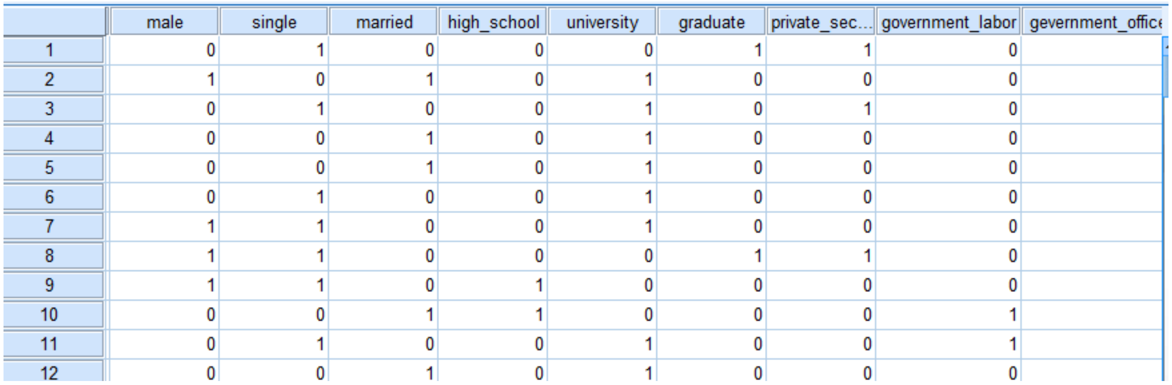
Although these results give us some information about the preference of Turkish investors, evaluating the results of descriptive analysis is the secondary aim of this study. The main purpose of this study is to focus on how clustering method can be used to segment the individual investors and based on this segmentation how the better decisions can be made regarding investment alternatives for both individuals and professional investors.

Preparing Dataset for Clustering Analysis

To cluster the investors based on socio-economic attributes, k-mode approach which is the extension of the k-means clustering approach is applied. Unfortunately, not all the software package even very commonly used ones are not readily available to conduct k-mode clustering algorithm. Therefore, some conversion is required to use k-means algorithm for clustering categorical data. This conversion can be made by translating each unique category to a dummy binary attribute as 0 and 1 to indicate the categorical value either absent or present in data record.

The software package used in this study is not capable of the making clustering directly for binary and categorical variables. Therefore, as to process the binary and categorical variables using software package, the values of the variables are translated into the appropriate format for preparing dataset for clustering analysis. The data acquired by conducting this transformation are shown in Table 4.

Table 4. Dummy Binary Variables of the Attributes



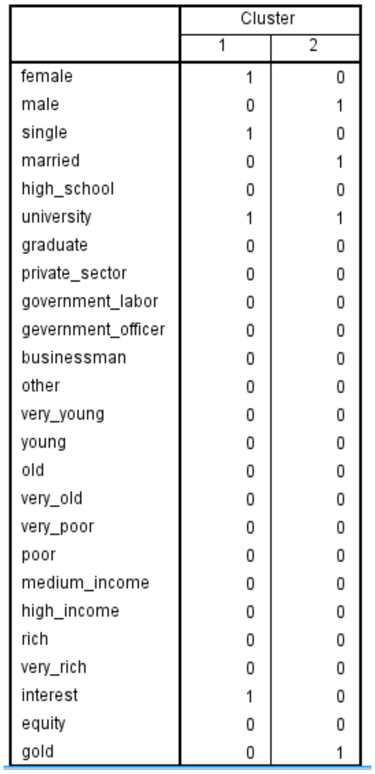
	male	single	married	high_school	university	graduate	private_sec...	government_labor	gevernment_office
1	0	1	0	0	0	1	1	0	
2	1	0	1	0	1	0	0	0	
3	0	1	0	0	1	0	1	0	
4	0	0	1	0	1	0	0	0	
5	0	0	1	0	1	0	0	0	
6	0	1	0	0	1	0	0	0	
7	1	1	0	0	1	0	0	0	
8	1	1	0	0	0	1	1	0	
9	1	1	0	1	0	0	0	0	
10	0	0	1	1	0	0	0	1	
11	0	1	0	0	1	0	0	1	
12	0	0	1	0	1	0	0	0	

As shown from Table 4, each answer obtained from responders is represented by 0 or 1 dummy variable indicating gender, marital status, education level, job etc. of the investors. In a dummy coding, each attribute is coded according to whether it belongs to the defined set or not. For instance, the first column represent gender and for gender male is coded as 1. Thus, if the respondent's gender is male then gender attribute value become 1 otherwise 0.

Clustering Analysis Result

Clustering model does not contain any outcome or dependent variable as stated before. It is an analysis method developed to create the clusters based on the similarity and distance relations of the data. In this study, not any predetermined outcome was defined, yet each investment option was considered as centroid of the cluster and other features were associated with them. Thus, it was aimed to determine the most appropriate investment alternative for investors whose socio-demographic characteristics were known. The k-mode technique, which is the extension of the k-means method, was used for this analysis based on 0 and 1 encoding. The results of the analysis obtained were shown in the Table 5.

Tablo 5. Clusters Based on the Investors' Socio-Demographic Attributes



	Cluster	
	1	2
female	1	0
male	0	1
single	1	0
married	0	1
high_school	0	0
university	1	1
graduate	0	0
private_sector	0	0
government_labor	0	0
gevernment_officer	0	0
businessman	0	0
other	0	0
very_young	0	0
young	0	0
old	0	0
very_old	0	0
very_poor	0	0
poor	0	0
medium_income	0	0
high_income	0	0
rich	0	0
very_rich	0	0
interest	1	0
equity	0	0
gold	0	1

As closely invesitigating results revealed some outcomes. The first one was that only two investment alternative emerged. It means that the third investment alternative, which must be stock option based on the frequency table, did not gather enough member to make cluster. Also, this result indicated that the investors of the stock option did not have very common attributes.

The second one was that some attributes have only 0 or 1 values for both clusters. It means that these characteristics were not effective in creating clustering. These attributes were age, income level and occupation. The other attributes took 0 and 1 value for different cluster. It means that they were effective attributes to distingusih two clusters. Namely, they were gender, marital status and education level. These findings are the same factors with studies made by Obuyami, 2013; Shafi, 2014; Dogan, Yildiz, & Topal, 2016. These outcomes could be more clearly visible by converting Table 5 results into Table 6 results.

Table.6. The Effective Factors (Attributes) Constructing Clusters

Cluster 1 (interest)	Female, single, university graduates
Cluster 2 (gold)	Male, married, university graduates

Table 6 demonstrated that the main characteristics of interest investors were women, single and university graduates. On the other hand, male, married and university graduates preferred gold investing. The result revealed that university graduates invested in both investment alternatives. The absence of stock cluster indicated that there were not enough investors to construct one cluster.

Conclusion

Making appropriate investment decisions based on the personal characteristics of individuals is very crucial for both individual and institutional investors. For these kinds of decisions, data mining techniques are applied recently. One of the most common data mining technique is clustering analysis. Therefore, the purpose of this study is to demonstrate the availability and applicability of cluster analysis in determining the best investment preferences that are the most suitable for investors based on their socio-demographic characteristics. In fact, this goal is similar purpose of clustering analysis in market segmentation. In the market segmentation, the aim is to develop appropriate strategies by taking into account the characteristics of the individuals to create clusters.

To create clusters, it is crucial to determine the characteristics of the individuals. It is known from practice that socio-demographic characteristics of individuals are vital and these attributes are effective in investment decisions as in many other individual decisions. The study aimed to reveal this relationship, which is thought to exist between these characteristics and investment choices applying clustering analysis. To achieve this aim with this approach, 332 individual investors were questioned via survey method using convenient sampling. However, 312 answers were applicable for analysis. The questions asked were regarding their investment preferences with their socio-demographic characteristics (attributes). Both socio-demographic characteristics and their investment preferences were categorical variables. Therefore, firstly data set were prepared for analysis converting these categorical variables into binary variables to conduct the clustering analysis. Then k-means algorithm which is available for numerical variables was conducted to get results. The results indicated that there have been only two clusters, because there were not enough stocks investors. Moreover, investors who invest in interest are women, single and university graduates, and male, married and university graduates investors preferred gold investing. Thus, the aim has been achieved illustrating these investors are risk averse and they prefer safe investment choices based on their socio demographic attributes.

This study can be extended in different ways based on dataset or method. For dataset, Likert-type questions can be developed by measuring investors' feeling and attitudes. These type questions can be analyzed applying fuzzy-clustering methods. For method, this algorithm can be conducted as an intermediate method using its outcomes as input values for other methods such as decision tree or logistic regression which are learning methods. In addition, due to the cluster analysis method as a machine learning process, this analysis can become part of the information systems and thus a new observation is added, the system can automatically rebuild all clusters each time.

The relationship between the investors characteristics and investment preference has been examined for long time in literature. However, almost all these studies applied classical statistical techniques. These techniques are not enough appropriate to respond quickly, effectively and accurately to changes in dataset. This study contributes to the literature introducing a clustering analysis as a new approach that can enable the process of investment decisions to be performed more effectively.

References

Akbulut, R., & Kaderli, Y. (2009). Sanliurfa İl Merkezindeki Borsa Yatirimcilarinin Profili ve Bu Yatirimcilarin Hisse Senetlerine Yatirim Yapma Surecini Etkileyen FaktOrlerin Analizi. *Muhasebe ve Finansman Dergisi, 43*, 212-226.

Aksoy, E. E. (2018). Turkiye'deki Bireysel Yatirimcilarin Yatirim Riski Tercihlerini Etkileyen Bireysel FaktOrlerin Analizi. *İsletme Arastirmalari Dergisi, 10*(3), 874-891.

Ali, I., & Rehman, K. U. (2013). Stock Selection Behaviour of Individual Equity Investors' in Pakistan. *Middle East Journal of Scientific Research, 15*(9), 1295-1300.

Angi, G. G., Bekci, I., & Karatas, O. N. (2016). Bireysel Hisse Senedi Yatirimcilarinin Bilissel Onyargilari Uzerine Bir Arastirma. Journal of Accounting & Finance. *Journal of Accounting and Finance, 70*, 171-192

Aydin, U., & Agan, B. (2016). Rasyonel Olmayan Kararlarin Finansal Yatirim Tercihleri Uzerindeki Etkisi: Davranissal Finans Cercevesinde Bir Uygulama. *Ekonomik ve Sosyal Arastirmalar Dergisi*, 95-112

Chauhan, N. S. (2020). *Introduction to Image Segmentation with K-Means clustering* . Retrieved from https://towardsdatascience.com/introduction-to-image-segmentation-with-k-means-clustering-83fd0a9e2fc3.

Collu, D. A. (2018). Hisse senedi Secim Karari Vermede Etkili Olan Bireysel Yatirimci Davranislari: BIST Ornegi. *Business and Economics Journal, 9*(3), 559-578.

Demir, Y., Akcakanat, T., & Songur, A. (2011). Yatirimcilarin Psikolojik Egilimleri ve Yatirimci Davranislari Arasindaki İliski: İMKB Hisse Senedi Yatirimcilari Uzerine Bir Uygulama. *Gaziantep Universitesi Sosyal Bilimler dergisi, 10*(1), 117-145.

Dogan, M., Yildiz, F., & Topal, Y. (2016). The change of investment preference by demographic characteristic: a survey on banking employess in Turkey. *Journal of Accounting, Finance and Auditing Studies, 2*(3).

Dolnicar, S. (2003). Using Cluster Analysis for Market Segmentation Typical Misconceptions, Established Methodological Weakness and Some Recommendations for Improvement. *Australasian Journal of Market Research, 11*(2), 5-12.

Faroog, A., Afzal, M. A., Sohail, N., & Sajid, M. (2015). Factors affecting investment decision making: Evidence from equity fund managers and individual investors in Pakistan. *Journal of Basic and Applied Scientific Research, 5*(8), 62-69.

Geetha, S. N., & Vimala, K. (2014). Perception of household individual investors towards selected financial investment avenues. *Procedia Economics and Finance, 11*, 360-374.

Goetzmann, W. N., & Kumar, A. (2008). Equity portfolio diversification. *Review of Finance, 12*(3), 433-463.

Gumus, F., Koc, M., & Agalarova, M. (2013). Bireysel Yatirimcilarin Yatirim Kararlari Uzerinde Etkili Olan Demografik ve Psikoljik FaktOrlerin Tespiti Uzerine Bir Calisma: Turkiye ve Azerbaycan Uygulamasi. *Kafkas Univeristiesi İktisadi ve İdari Bilimler Fakultesi Dergisi, 4*(6), 71-93.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* Elsevier.

Hierarchical Clustering / Dendrogram: Simple Definition, Examples. (2020, April 15). Retrieved from Statistics How to: https://www.statisticshowto.com/hierarchical-clustering/

Huang , J. Z. (2009). Clustering Categorical Data with K-modes. In *Encyclopedia of Data Warehousing and Mining* (p. 5). IGI Global .

Huang, J. Z. (2009). Clustering Categorical Data with k-Modes. IGI Global.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery, 2*, 283-304.

İslamoglu, M., Apan, M., & Ayvali, A. (2015). Determination of factors affecting individual investor behaviours: A study on banker. *International Journal of Economics and Financial Issues, 5*(2), 531-543.

Kapoor, S., & Prosad, J. M. (2017). Behavioural finance: A review, procedia computer science. *122*, pp. 50-54.

Khan, S. S., & Amir, A. (2013). Cluster center initialization algorithm for K-modes clustering. *Expert Systems with Applications*.

Lodhi, S. (2014). Factors Influencing Individual Investor Behaviour: An Emprical Study of City Karachi. *IOSR Journal of Business and Management, 16*(2), 68-76.

Obuyami, T. M. (2013). Factors Influencing Investment Decisions on Capital MArket: a study of Individual Investors in Nigeria. *Organizations and Markets in Emerging Economics 4, 1*(7), 141-161.

Patlola, C. R. (2020). *Understanding the concept of hierarchical clustering technique.* Retrieved from towards data science: https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

Rezankova, H., Praze, S. E., & Vysoka, P. (2009). Cluster analysis and categorical data. *Statistika 89*.

Shafi, M. (2014). Determinants Influencing Individual Investor Behaviour in Stock Market: a Cross Country Research Survey. *Arabian Journal of Business and Management Review, 2*(1), 60-71.

Sunden, A. E., & Surette, B. J. (1998). Gender differences in the allocation of assets in retirement savings plans. *The American Economic Review, 88*(2), 207-211.

Tuma, M. N., Decker, R., & Scholz, S. W. (2011). A Survey of the Challenges and Pitfalls of Cluster Analysis Applicaiton in Market Segmentation. *International Journal of Market Research, 53*(3), 391-414.

Usul, H., Bekci, I., & Eroglu, A. H. (2002). Bireysel Yatirimcilarin Hisse Senedi Edimine Etki Eden Sosyo-Ekonomik Etkenler. *Erciyes Universitesi İktisadi ve İdari Bilimler Fakultesi Dergisi, 19*, 135-150.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making.* Milano: Wiley&Sons, Ltd. .

Wameru, N., Munyoki, M., & Uliana, E. (2008). The effects of behavioral factors in investment decision making: a survey of institutional investors operating at the Nairobi Stock Exchange. *International Journal of Business and Emerging Markets, 1*(1), 24-41.

Yesildag, E., & Ozen, E. (2015). Usak İlindeki hisse senedi yatirimcilarinin profili ve yatirim kararlarini etkileyen demografik ve sosyo-ekonomik faktOrlerin analizi. *Journal of Accounting, Finance and Auditing Studies, 1*(2), 78-102.