Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

Gulzhaukhar Assanbayeva

Principal component analysis and it's application to analysing factors affecting choice of university major

THESIS
Presented in Partial Fulfillment for the
Degree of Master of Science in Mathematics
(degree code: 6M060100)
Department of Natural sciences
Faculty of Engineering and Natural Sciences

Supervisor: **Shirali Kadyrov**
Kaskelen, 2020

Abstract

The study is considered the identification of the main factors affecting the selection of study field.As an application, we develop a construct to measure factors that affect college students in their major selection.It is a multilingual building in three languages,namely,Kazakh,Russian,English.We conducted a survey of 27 Likert scale items in 3 languages,which is being conducted among 314 undergraduate students in Kazakhstan.The overall reliability of the test is calculated to be 0,856.The nine scales are the effect of Uniform National Testing, state grant affect, personal interest affect, skills affect, occupation salary affect, teacher affect, external affect, university cost affect, parent's affect.The result of study can be used to help universities to improve their competences in attracting students.For these used PCA method and as a software used a Python programming language. Also,we investigate the main factors that influenced the Kazakhstani students most.We tested any differences in terms of gender when it comes to major selection factors.As most participants were among Suleyman Demirel University,we also wanted to see if there any differences between factors that influence Suleyman Demirel University students major selection to students of other local universities.The statistical data analysis show that students found themselves to be the most influential by 6 factors out of 9 ,namely,influence of state grant, influence of university cost, occupation salary impact, national test impact, influence of personal interest and influence of personal skills.Hypotheses testing are carried out using the T-test.Findings also suggested that, there is no difference between these 6 factors depending on gender.Also,no differences were found among university and factors.

Аңдатпа

Бұл зерттеуде біз әртүрлі қолданбалы салаларда енгізілген сызықтық алгебрадан белгілі негізгі компонентті талдаудың артындағы математиканы қарастырамыз.Бағдарлама ретінде біз университет студенттеріне негізгі мамандық таңдау кезінде әсер ететін факторларды өлшейтін сауалнама жасаймыз.Бұл үш тілде, атап айтқанда қазақ, орыс және ағылшын тілдерінде берілген көптілді сауалнама. Осы мақсатта авторлар үш тілде 27 Likert шкаласынан тұратын сауалнама дайындады және ол Қазақстандағы 314 студенттер арасында өткізілді.Өлшемділікті төмендету үшін негізгі компоненттік талдау Python арқылы есептелінді, нәтижесінде 22 негізгі элементтерден тұратын 9компоненттер алынды.Тесттің жалпы сенімділігі 0,856 құрайды.Тоғыз шкалалар: ұлттық тестілеу нәтижесі әсері, мемлекеттік грант нәтижесі,жеке қызығушылық әсері ,өз қабілетінің әсері, мамандықтың жалақысы әсері мұғалімнің әсері, сыртқы әсер, университеттің құнынының әсері,ата-ананың әсері.Сондай ақ біз қазақстандық студенттерге көп әсер еткен негізгі факторларды зерттейміз.Біз іріктеудің басты факторлары туралы сөз болғанда, жыныс тұрғысынан кез келген айырмашылықтарды тексердік.Қатысушылардың көпшілігі Сулейман Демирел университетінің студенттері арасынан болғандықтан, басқа жергілікті университеттер студенттерінің арасында Сулейман Демирел университетінің студенттерінің негізгі іріктеуіне әсер ететін факторлар арасындағы қандай да бір айырмашылықтар болғанын тексердік. Статистикалық мәліметтерді талдау студенттер 9 фактордың 6 бойынша аса ықпалды болғанын көрсетеді,атап айтқанда:мемлекеттік гранттың әсері,университет құнының әсері,жалақы мамандығы- ның әсері, ұлттық тестінің әсері, жеке қызығушылықтың әсері және жеке дағдылардың әсері.Алынған нәтижелер, сондай-ақ осы 6 факторлар арасында жынысына байланысты ешқандай айырмашылық жоқ екенін куәландырады.Сонымен қатар, университет пен факторлар арасында ешқандай айырмашылық жоқ.

Аннотация

В данном тезисе мы рассматриваем математику,лежащую в основе хорошо известного анализа главных компонентов из линейной алгебры, реализуемой в различных прикладных областях.В качестве приложения мы разрабатываем конструкцию для измерения факторов,влияющих на студентов университета в их главном выборе.Это многоязычная конструкция, представленная на трех языках, а именно на казахском, русском и английском.С этой целью авторы подготовили опрос, состоящий из 27 пунктов шкалы Лайкерта на трех языках,и он был проведен среди 314 студентов бакалавриата Казахстана. Для уменьшения размерности был проведен анализ главных компонент в python,который привел к 9 основным масштабам с только 22 элементами. Общая достоверность испытания, по расчетам, составляет 0,856. Девять шкал:влияние единого национального тестирования,влияние личного интереса,влияние государственного гранта,влияние заработной платы по профессии,влияние навыков, влияние преподавателя,влияние внешних факторов,влияние стоимости университета, влияние родителей.Также мы исследуем основные факторы,которые больше всего повлияли на казахстанских студентов.Мы проверили любые различия с точки зрения пола, когда речь заходит о главных факторах отбора.Поскольку большинство участников были из числа студентов Университета Сулеймана Демиреля,мы также хотели посмотреть,есть ли какие-либо различия между факторами,влияющими на основной отбор студентов Университета Сулеймана Демиреля среди студентов других местных университетов.Анализ статистических данных показывает,что студенты оказались наиболее влиятельными по 6 факторам из 9,а именно:влияние государственного гранта,влияние стоимости университета,влияние заработной платы по профессии, влияние Национального теста, влияние личного интереса и влияние личных навыков.Полученные результаты также свидетельствуют о том, что между этими 6 факторами нет никакой разницы в зависимости от пола.Кроме того, не было обнаружено никаких различий между университетом и факторами.

Contents

1. Introduction

The selection of study field is a very important question that should be dealt with carefully.First, your personal interests need to be considered. Second, assess their own abilities. Thirdly, to analyze the correspondence of their personal qualities with those qualities, the presence of which requires the chosen profession. Fourth, to study the possible life prospects that will give this or that profession.Dreams can be purely fictional. In order to realize your dreams you need to know more information [5].The forefathers of mankind sought to ensure the moral freedom of the individual. Making decisions is the logical way to set your mind to choose from among many other people [3]. The ability to satisfy the ease of man. This is true, as it is said that good decision making is a necessary skill for career success in general and effective leadership in particular. It's true that in order for a person to be successful, he must have good decision making.Every day we make decisions, big or small, and these decisions can make an impact on our life.After graduating from high school children face with their most important decisions and that's the future career decision. Students are faced with the fact that they will be resolved in the future by this particular field decision. The alternative is to help the child in which area to study. Making your own decision helps a person develop and advance in life.Choosing a selection field of study is vital not solely in one's educational life [25]. However, additionally within the future personal life as a result of it's an effect on the standard of education, student satisfaction, career and job opportunities, money compensation and at last social standing. However, this choice of specialty can be a nervous and oppressive task as a result of the fact that students do not create this challenge in a vacuum; There are various factors that affect this decision. They are based on the very fact of choice, increase student satisfaction.

1.1 Literature review

A study of the general social science literature indicates two key work emphases in the field of major selection.First, and more specifically relevant to existing work, there is a general body of information that describes particular variables that affect the option of main choices.Secondly, there is a branch of general analysis that reflects on the interaction between different factors (e.g. ethnicity and gender and academic year) and the primary preference. There are various questionnaires are used in the literature to analyse factors related to major selection. Factors such as Interest in major, Peer pressure, Family pressure, Academic ability, Major's reputation, Job availability, Job salary, Major's prestige, Public sector job, Private sector job were analyzed in Aldosary [1] from 447 students of King Fahd university and Job availability, Salary, social status and prestige were found to be the main affecting factors in that order. Another study was carried with 111 participants to investigate college students' academic major declaration [9]. An exploratory factor analysis was carried by [25], to analyse the variables affecting the specialization selection of 300 business graduates in Lahore resulting 6 main factors: academic factors, social capital factors, future prospect factors, human capital factors, market demand factors and finally job prospect factors. This 31-item construct is calculated to have high reliability of 0.845. Another study was carried [8] at the University of Tennessee, Martin to determine the variables that influence agriculture students' choices in deciding their career path. The findings show that the main variable (22%) is family influence followed by a factor "a career that is personally rewarding" (21%).Taylor noticed that students had proclaimed their curiosity to be the most significant factor to pick a class.Curiosity indicated in the area has taken a more successful position for educational students than for business students in [7].In a national analysis by [22], it was observed that although students in the United States considered internal motivations to be the most important factor in the selection of teaching profession, students in Cyprus decided external influences to be the greatest force. Second, and similar to existing studies, it reflects on the interaction between demographic data and the items that influence selection of specialization.Study in this area relies on a number of outlets.The study has found that students who select business major are influenced by parents and socio-economic class,with the extent of this

impact differing according to gender[13].[25]observed by six factors that the effect of influences on the selection of specialization between males and females was the same.It was found that there was no relationship between the universities and the main factors.[17] observed that women appeared to assign more priority to performance than men.However, [6] there was no demographic gap in the weighting of accounting in terms of family support and advice from colleagues.

Linear algebra is a branch of mathematics which deals with the system of linear equations,vector spaces, linear maps and their properties. Matrices are one of the building blocks of linear algebra. A numerical data consisting of cases$m$ and n variable entries for each case can be thought of as matrix. This representation enables us to carry various manipulations available to us from linear algebra and interpret the results. If n is large, it often becomes difficult to derive meaningful conclusions from the data. Principal Component Analysis (PCA) is one of the widely used techniques from linear algebra that helps with dimensionality reduction and makes it possible to extract hidden features of the data [24]. Even though this is a century old method invented by K. Pearson ([23]), in its original form and in improved versions, is still being used nowadays for handling various large datasets. Some of the research areas where PCA is used include signal processing [30], genetics[14], quantitative finance[2], neuroscience[28], and questionnaire development [4].A PCA method applied to determine the main factors that influences choice of major.The works we have mentioned above,offer proof of knowing the main variables that affect the selection of major.

1.2 Research objectives

1. To develop a multilingual construct that measures factors in university major selection using Principal Component Analysis

2. Use this construct to study the main factors that influence university major selection in Kazakhstan.
   a)To master in PCA method from Linear Algebra.
   b)To excel in data analysis using Python programming languages.
   c)To determine the most influential factors for Kazakhstani students.
   d)To find relationship between main factors and demographic data.

2. Methods

2.1 Principal component analysis

Principal component analysis (PCA), is one of the main methods used to reduce the dimension of data with the least loss of information. Developed by in 1901 by Karl Pearson is Used in image recognition, data compression, computer vision, etc. the Calculation of the main components is reduced to the calculation of eigenvectors and eigenvalues of the matrix of the original data. The principal component method is also known as the Karhunen-Loeve transform or the Hotelling transform(KLT)[23]. From a mathematical point of view, the principal component method is an orthogonal linear transformation that maps data from the original feature space to a new space of smaller dimension.In this case, the first axis of the new coordinate system is constructed in such a way that the data dispersion along it would be maximal. The second axis is built orthogonally to the first one so that the data dispersion along it is the maximum of their remaining possible values, and so on[15]. The first axis is called the first main component, the second - the second, and so on.Therefore, the task of the principal component method is to construct a new feature space of a smaller dimension, the variance between the axes of which will be redistributed so as to maximize the variance for each of them. To do this, follow these steps[15]:

1. The total variance of the original feature space is calculated. This cannot be done by simply summing the variances for each variable, since they are not independent in most cases. Therefore, we need to sum up the mutual variances of variables that found from the covariance matrix.

2. Eigenvectors and eigenvalues of the covariance matrix are calculated that determine the directions of the principal components and the amount of variance associated with them.

3. The dimension is reduced. The diagonal elements of the covariance matrix show the variance from the original coordinate system, and its eigenvalues from the new one. Then dividing the variance associated with each main component by the sum of the variances for all components, we get the proportion of variance associated with each component. After this, so many main components are discarded that the percentage of remaining components is 80-90%.

The dimension is reduced. The diagonal elements of the covariance matrix show the variance in the original coordinate system, and it should be noted that the directive approach to choosing the number of components does not always give good results. This is due to the fact that part of the data variance may be due to noise, rather than the information content of components. Then, if you set a threshold of, say, 80%, you might find that only 60% of the variance is related to information content, and 20% is related to noise. Therefore, in practice, various special criteria are often used to define the number of components, such as the Kaiser criterion, the broken cane criterion, and so on.

The main limitations of the principal component method are:

1. It is impossible to interpret the components meaningfully, since they "absorb"
   the variance from several source variables;

2. The method can only work with continuous data.

The principal component analysis is included in most analytical platforms and is widely used to decrease the dimensionality of input data at the preprocessing stage.The method is sometimes considered as part of a more General approach to reducing the dimension of data - factor analysis. In analytical platforms, it is often the principal component analysis that is practically implemented in factor analysis modules.

2.1.1 Linear algebra behind PCA

A PCA is an application of linear algebra where one rotates and shifts the coordinate axes to obtain more suitable representation of data helpful for feature

extraction,one that presents important information. PCA requires a small background of linear algebra. So, we now discuss some basic concepts of linear algebra, in particular algebra [12] used to apply in PCA. Eigenvectors and eigenvalues are important properties of matrices which are necessary for PCA[28].

Definition 2.1. Let A Let be a real matrix. A complex number is called an eigenvalue of a matrix if there exists a dimensional non-zero complex vector , called an eigenvector, such that

$$A\vec{x} = \lambda\vec{x} \tag{2.1}$$

To determine eigenvalues,one needs to solve the characteristic equation:

$$D(\lambda) = det(A - \lambda I) \tag{2.2}$$

By solving the equation for λ, we will have eigenvalues $\lambda_1, \lambda_2...$ .By substituting λinto the vector equation, we can obtain eigenvectors.

Eigenvectors belonging to different eigenvalues are easily seen to be linearly independent. If a matrix is symmetric, then in fact distinct eigenvectors are mutually orthogonal[12]. We now make these notions more clearer. Orthogonality is significant since it ensures that the data can be represented in terms of such perpendicular eigenvectors, rather than interpreted in terms of the x and y axes[28].

Definition 2.2. An $m \times n$ matrix $A = [a_1, a_2...a_n]$ is said to be orthogonal

$$a_i^T a_j = \begin{cases} 1 & \text{if } i = j \\ 0 & otherwise \end{cases} \tag{2.3}$$

where each a_i,i=1,2,3...,n is a column vector of m in rows

Theorem 2.3. *The inverse of an orthogonal matrix is that it's being transposed[12].*

Definition 2.4. A $m \times n$ square matrix A is said to be symmetric if A_{ij}=A_{ji}, i.e., row index and column index are interchangeble: $A^T = A$

Theorem 2.5. *For any $m \times m$ matrix of real numbers A, $m \times m$ matrix $A^T A$ and the $n \times n$ matrix AA^T are symmetric[12]*

Proof: Let's take the transpose AA^T. We apply properties of transpose operation.

Then:

$$(AA^T)^T = A^{TT}A^T = AA^T$$

we repeat this analysis for A^TA

$$(A^TA)^T = A^TA^{TT} = A^TA$$

Definition 2.6. A matrix A is said to be diagonalizable if there exists some B such that $A = BDB^T$, where D is a diagonal matrix and B is some special matrix that diagonalizes A. Additionally, if B is orthogonal, then B is said to be orthogonally diagonalizable [28].

Theorem 2.7. *The matrix becomes symmetric if it is orthogonally diagonalizable[12].*

Proof: Suppose A is orthogonally diagonalizable.Let us compute A^T.

$$A^T = (BDB^T)^T = B^{TT}D^TB^T = BDB^T = A$$

Hence,if A is orthogonally diagonalizable,it must also be symmetric.

Theorem 2.8. *If A is symmetric(meaning$A^T = A$),then Ais orthogonally diagonalizable and has only real eigenvalues.In other words,there exist real numbers $\lambda_1, \lambda_2...\lambda_n$ (the eigenvalues) and orthogonal,non-zero vectors $\{\vec{v_1}..\vec{v_n}\}$ (the eigenvectors) such that for each i=1,2...n[12].*

$$A\vec{v_i} = \lambda_i\vec{v_i}$$

[28] Let A be square n x n symmetric matrix with associated eigenvectors $e_{i\,i=1}^{n}$ and $E = [e_1.....e_n]$.Then:

Theorem 2.9. *A symmetric matrix A is diagonalized by a matrix of its orthonormal eigenvectors[12].*

Proof: This theorem establishes that the diagonal matrix D operates in such a way that $A = BDB^T$.Let A be some matrix, not inherently symmetric, and let it have independent eigenvectors.

$$AB = [Ae_1.....Ae_n] = [\lambda_1e_1.....\lambda_nn] = BD$$

Since AB=BD,it follows that $A = BDB^{-1}$.

2.1.2 Statistics behind PCA

Standard deviation and variance

Mean is another average name. To define the mean of the data collection, add all values and divide the amount of total values in the sample. We will find that with the following formula:

$$\overline{X} = \frac{\sum X}{n}$$

Standard deviation is a calculation of the dispersion from its mean of a set of data.The larger the dispersion or variation, the higher the standard deviation and the greater the meaning divergence from its norm.

$$s = \frac{\sum (X - \overline{X})^2}{N-1}$$

Mean gives information about the middle point,but it does not say too much about data.

Statistical variation (σ^2) is a calculation of the difference between numbers in a data set. That is, it calculates how far an increasing number in the system is from the average, and thus from any other number in the set.

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{N-1}$$

where: x_i =the i-th data;
$\overline{X}$=the total mean;
n=the number of data;

Covariance

Standard seviation can only be found independently of other measurements and only works on one axis.A measure like this is covariance[28].

Covariance is often measured between two dimensions.We get the discrepancy when you evaluate the covariance between the variable and the entity itself.If you had a set of 3-dimensional data*(x, y, z)*, so you might calculate the covariance between the dimensions x and z, the dimensions x and y, and the dimensions y and z.Covariance indicates how two factors are connected to each other.Positive covariance implies that the variables are positively connected, whereas negative covariance indicates that the variables are opposite.

The formula for determining the covariance of data is given below[28].

$$var(X) = \frac{\sum (X_i - \overline{X})(X_i - \overline{X})}{n-1}$$

Same as variance formula,covariance will represent next formula:

$$Cov(X, Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

n = number of data points in the sample
$\overline{X}$= the mean of the independent variable x
$\overline{Y}$= the mean of the dependent variable y x = the independent variable
y = the dependent variable

You can use more than 3 dimensional data sets (x, y, z), you can measure *cov*(x, y), *cov*(y, z),*cov*(x, z).We can assume that you can consider that for the n-dimensional data collection [12] $\frac{n!}{(n-2)!*2}$ different covariance values. If the covariance value is positive, this indicates that the both dimensions increase together.If the value is negative, then if one factor decreases, the other will increase.In the end, if the covariance is 0, it indicates that the two dimensions are independent of each other. By using covariance ,you can construct square matrix,$C_{i,j} = \sigma(x_i, x_j)$ where $C \in R^{d \times d}$ d is a number of random variables of the data.The covariance matrix is symmetric $\sigma(x_i, x_j) = \sigma(x_j, x_i)$.The calculation for the covariance matrix can be also expressed as follows[28]:

$$C = \frac{1}{n-1} \sum (X_i - \overline{X})(X_i - \overline{X})^T$$

In a 3 dimensional data set x,y,z, you represent matrix as follows:

$$\begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

2.1.3 Step by step calculation of PCA

Step 1: Preparation of data As an example, we're going to use the set data. We should demonstrate what the analysis of the PCA is doing at every stage.[28]

Step 2: Standardize

In order for PCA to operate correctly, we need to remove the mean from each of

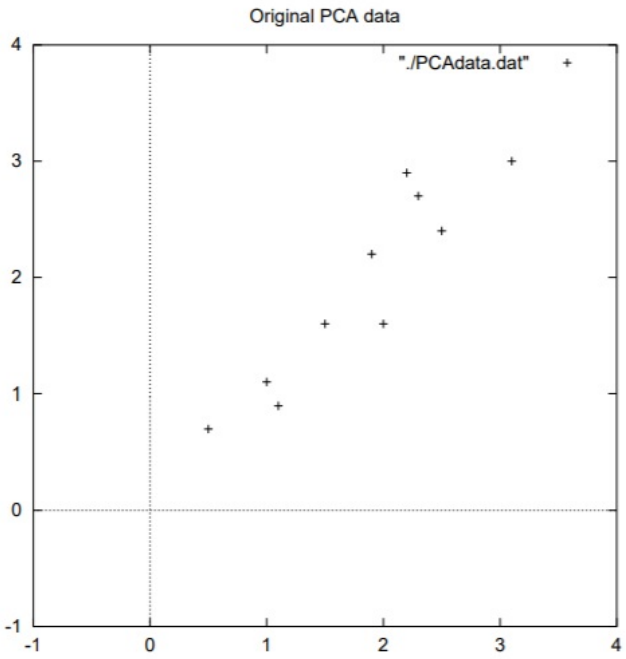
the data dimensions.It helps to centralize our data. So, all the meanings for $\overline{x}$ (average of x values over all data points) subtracted, other all the y values have $\overline{y}$ subtracted from them. It generates a data with a mean of 0[28].

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01



Step 3: Calculate the matrix of the covariance

Because the data is 2 dimensional, the covariance matrix should be 2x2. There are no surprises here, and we're only going to send you the result with 3 d.p[28]:

$$cov = \begin{pmatrix} 0.617 & 0.615 \\ 0.615 & 0.717 \end{pmatrix}$$

Since the non-diagonal elements in this covariance matrix are positive, both the x and the y variables may be assumed to increase together..

Step 4:Find the eigenvectors and eigenvalues of the covariance matrix[1]

Covariance matrix is a square matrix,then we may determine for this matrix,the eigenvector and its eigenvalue. This is very significant, because they provide us with useful data knowledge. Eigenvectors and Eigenvale currently[28]:

$$eigenvalues = \begin{pmatrix} 0.0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{pmatrix}$$

It should be noted that eigenvector both are unit eigenvectors, it means that their length is one. This is important for the main component analysis, but luckily, most math packages, when you ask for their eigenvectors, send you the unit eigenvector.We evolve along with two variables, as predicted from the covariance matrix. We created eigenvectors at the top of the data. In the tale they given as diagonal dotted line. They are perpendicular to each other, as mentioned in the segment of their eigenvectors. But most importantly, they give us data information on the laws.Where is one of the eigenvectors going around the midpoints, like drawing a line with the best fit?The eigenvector tells us how these two data sets bind to that axis.The 2nd eigenvector provides us with another non-essential illustration of the reality that all other points move on the main line, but to some degree on the main line side.Thus,we can develop the lines, which characterize the data as a result of the covariance eigenvector matrix.Rest steps are to convert the data, so that they are displayed as lines.

Step 5: Selecting components and forming a feature vector

This is where the concept of compression and reduced dimensions of data enters.In the previous section, if you find your eigenvectors and beliefs, you may note that eigenvalue is a completely different concept.The eigenvector with maximum eigenvaluer is the main principal component of the data set.The patented eigenvector of its eigenvalues was the one leading down the data center.It is the most important connection between the dimensions of the data.Overall, after the identification of the eigenvectors from the convariance matrix, the next step to put them descending order of eigenvalues.It brings you the necessary components in sequence.And, if you like it, you could want to disregard the less essential elements.If any elements are removed, the final data set may have less dimensions than the original version.To be precise, if you initially have n dimensions in your results, and then calculate n of your eigenvectors and eigenvalues, and then pick

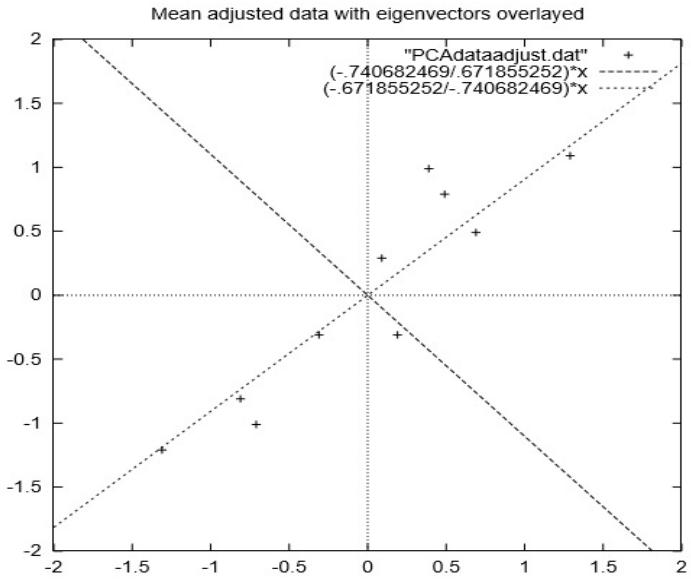


Figure 2.1: Normalised data plot

the highest p of your eigenvectors, so the final data set contains just p dimensions.What you need to do now is create a feature vector, which is just a nice word for a vector matrix. That is achieved by taking the eigenvectors you choose to avoid from the list of eigenvectors and then constructing a matrix in the columns of such proprietary vectors.

$$Featurevector = \begin{pmatrix} eig_1 & eig_2 & eig_3 ..... eig_n \end{pmatrix}$$

We will get 2 eigenvectors from given example.Then we can construct a feature vector with 2 of the eigenvectors with 3 d.p:

$$\begin{pmatrix} -0.678 & -0.735 \\ -0.735 & 0.678 \end{pmatrix}$$

We may opt to delete a less necessary component and only have one column with 3 d.p:

$$\begin{pmatrix} -0.678 \\ -0.735 \end{pmatrix}$$

Step 6: Constructing the new data set[28]

This is the last step in PCA, and the simplest one too. After choosing the components that we would like to use in our data and building a feature vector, we essentially take the transposal vector and subtract it to the left of the original

data collection.

$$Featuredata = RowFeatureVector \times RowAdjustData$$

Where the Row feature Data is the matrix with the proprietary vectors in the columns, such that the proprietary vectors are all in the rows, with the most important proprietary vector at the center, and the Row Data Change is the transposed mean-adjusted value. Final data is a complete data collection, with data elements in columns and measurements in line.It sends us the original data only in terms of the vectors we have selected. Our initial data collection had two axes, x and y, and our data was focused on them.It is possible to transmit data on any two axis you wish. If such axes are perpendicular, the expression is the most effective.That's why it was necessary that the eigenvectors should always be perpendicular to each other. We've modified our results from being in terms of the x and y axes, and now they're in terms of our two eigenvectors.In the case of a new data set, the dimensionality has been reduced. We've thrown some of the existing vectors out, the latest data is just in terms of vectors that we've chosen to hold.To demonstrate this on our results, we have rendered the final transformation with the possible vector features. In each scenario, we took the transposal of the outcome to get the results back to a comfortable table-like format.

2.2 Data collection

Data is the details you gather for the intent of addressing your study request. The types of analysis you choose rely on the sort of data you need.Data collection is a method of collecting and analyzing information on variables of interest in a systematic way that allows us to answer existing research issues, to test hypotheses and to review outcomes.The data collection element is important for all fields of research, including physical and social sciences, the humanities, business, etc. While the methods vary by practice, the emphasis on ensuring fair and equitable selection stays the same.Irrespective of the area of analysis or choice for identifying data (quantitative, qualitative), reliable data collection is important for preserving the credibility of the work. Both the availability of suitable data collection methods (existing, updated or freshly developed) and specifically specified guidelines for their proper usage minimize the risk of errors occurring[31].

2.2.1 Qualitative and quantitive data

Quantitative data may be counted, calculated and represented in numbers. Qualitative evidence is both informative and analytical. Qualitative data may be classified on the basis of features and attributes[32].
Qualitative data:These statistics are not usually calculated using hard numbers used to create graphs and maps. It is defined then on the basis of assets, characteristics, names, and other identifiers.Qualitative evidence may be used to determine why. This is questioned and is always open-ended before more work is performed. Generating these evidence from qualitative studies is used for explanations, definitions, observations and original understandings.
In comparison to qualitative data, quantitative data is descriptive and usually organized in nature – implying that it is more linear and established. This type of data is calculated using numbers and values, which renders it a more fitting candidate for data analysis.Though qualitative data is available for experimentation, quantitative data is far more descriptive and succinct. This may be used to pose "how many" accompanied by definitive detail[20].
Qualitative data may almost often be called unstructured or semi-structured data. This sort of data is configured dynamically with very little structure. Therefore, qualitative evidence can not be obtained and interpreted using traditional approaches.
Quantitative data can almost always be viewed as organized data. Each sort of data is structured in such a manner that it can be easily arranged and searchable in relational databases. The most growing definition of organized data might be the numbers and values used in spreadsheets.

2.2.2 Primary data

Primary evidence is all original knowledge you obtain for the intent of addressing your study question (e.g. by polls, analyses and experiments).If you're looking at a new study issue, you're definitely going to need to gather primary data.Primary facts are proof gathered by a researcher from first-hand knowledge, utilizing methods such as surveys, interviews or studies[31]. In the case of the research project, information is derived directly from primary sources.Primary data might be gathered to address the particular study request. You have power over methods of

measuring and analysis.However,primary is the less difficult and time intensive to process.Needs guidance on data processing data.

2.2.3 Survey

Survey is used as a data gathering tool in a number of areas. We are a perfect option if you decide to find out regarding the features, tastes, views or values of a community of individuals.Before beginning a survey, you will already have a simple study query that determines what you want to find out. Before you begin, establish a detailed strategy for where, where, how, and with whom to perform a survey. Define in advance how many of the responses you need and how you can get access to the study.On the basis of this issue, you need to decide precisely how you want to engage in the survey.There are two primary forms of surveys[21]:

1. A questionnaire, where the collection of questions is provided via fax, electronically or in person, and is filled out by the respondents themselves.A questionnaire is a list of questions that are used to collect details about someone or something. This is not intended for mathematical research or for identifying trends and patterns. An analogy may be whether you sign up for a workout or go to a check-up and have to address a series of questions regarding the present medical health.

2. Interview where the interviewer poses a series of questions by phone or in person and documents the replies. It is carried out for a intent such as study, analysis, and the like, where all parties engage in one to one contact.This is known to be one of the better forms of data collecting as it provides two avenues of sharing details, the interviewer gets to know the respondent, and the respondent knows about the interviewer.

Then, once you are confident that you have developed a good research design that is conducive to addressing your research questions, you can perform a survey utilizing your preferred form – via fax, online, or in person.There are a variety of ways to evaluate the findings of the survey. You have to interpret the data first, typically with the aid of a computer system to figure out all the responses. You can also clean up the details by deleting missing or wrongly executed answers.When you conducted open-ended questions, you would have to code the answers by

assigning marks to each response and grouping them into groups or themes. You may also use more comprehensive approaches, such as thematic analysis, which is especially useful for the study of interviews.

2.3 Sampling

Population: The whole community that you want to draw conclusions about is the nation.

Sample is the specific group of people you gather data from.

The selective method (method of sampling) is a statistical approach for analyzing the general properties of the entirety of some entity dependent on the analysis of the properties of just a portion of such objects. The entirety of the items examined which are of concern to the study is considered the general population.And some of the objects to be studied are called either a sample population or a sample[18]. The key purpose of the sample survey is to achieve the most reliable definition of the population of interest on the basis of the minimum sample size results. This can only be done on the basis of a representative survey, i.e. surveys with statistically measurable features of the general population.

The accuracy of the results of sample surveys is achieved through the use of sophisticated sampling methods (cluster selection, stratification, the use of probability-proportional selection,the simple random or random selection, repeated or non-repeated selection). The required sample size depends on several parameters of the analysis (the approximate indicator or set of indicators, the type and methods of sampling, the variance of the data tested, the stated precision of the tests, the maximum allowable error in the assessment of indicators) and is calculated on the basis of quantitative statistical calculations or by an expert.

The limited approach is used mainly in economics, marketing and clinical science. In addition, however, in the statistical analysis of data in either area, the researcher operates, as a rule, not with the general public, but with the sample.

Non-probability sampling is defined as a sampling strategy in which the researcher chooses samples on the basis of the researcher's individual opinion rather than a random selection.This is less strict form.The process of sampling relies highly on the skill of the researchers. It is carried out through evaluation and is commonly utilized through academics for qualitative studies.Non-probability

sampling is a sampling process in which not all values of the sample have the same likelihood of engaging in the survey, as opposed to probability sampling, where each member of the community has a defined risk of being chosen.Non-probability sampling is particularly effective for explotorious experiments (deployment of the test to a limited sample relative to the predetermined sample size).Scientists apply this approach in experiments where it is not feasible to perform random samples on the grounds of time or expense considerations[27].

Advantages of nonprobability sampling:

1. Non-probability sampling is a more effective and practical way for research to carry out surveys in the modern world. Although statistics like sampling odds, they produce results in the form of numbers. However, if performed correctly, the non-probability of sampling the produce comparable outcomes, if not the same consistency.

2. Response to non-likelihood sampling is quicker and more cost-effective compared to chance sampling as the question is described by the researcher. Respondents reply rapidly relative to randomly chosen participants who have a large degree of desire to participate.

Convenience sampling is a non-probability sampling process where samples are collected from the population precisely because they are readily available to the researcher.These samples are selected by researchers mainly because they are simple to identify and the analysis did not try to pick a group that represents the whole population.Ideally, in research, it is important to test a sample that represents the population. Nonetheless, among the different research, the population is too broad to assess and measure the society as a whole.It is one of the explanations why researchers focus on convenient sampling, the most popular non-probability sampling process, because of its speed, cost-effectiveness and ease of availability of the sample[27].

3. Instruments and tools

3.1 Measurement validity and reliability

Reliability and validity are criteria used to determine the consistency of study.We mean how good a process,procedure or examination tests something. Reliability is about the quality of the test,and its validity is about the precision of the calculation.It is crucial to remember reliability and validity when designing your study,preparing your methodology, and writing your findings, particularly in quantitative analysis.Reliability applies to how reliable a system is to calculate it.If the same outcome may be reliably obtained by using the same methodology in the same conditions, the calculation is assumed to be accurate. Validity relates to how specifically a system calculates what it is supposed to calculate. If research is of good standard, this means that it offers results that apply to real properties, characteristics and discrepancies in the physical or social setting.High precision is one predictor of the quality of the calculation.If a procedure is not accurate, it will not be true[16].

3.1.1 Reliability test

Cronbach's alpha (or alpha coefficient),invented by Lee Cronbach in 1951, tests reliability or internal accuracy."Reliability" is how accurately the check tests what it will do[16].For starters, an organization can provide its workers with a work satisfaction survey.High reliability implies that it tests work satisfaction, whereas low reliability suggests that it tests something different (or even none at all).Cronbach's alpha searches to see if multiple-question Likert scale polls are correct.Such questions test implicit variables — hidden or unobservable variables such as: conscientiousness, neurosis, or transparency. They are very challenging to measure in actual life[16].Cronbach's alpha will inform you whether the check

you have developed is to calculate the element of concern reliability.Shows the internal accuracy of the attributes identified by the object.It is also used in psychology to evaluate the efficacy of the psychological evaluation.We will use the Greek letter α.Calculated using the formula:

$$\alpha = \frac{N * r}{1 + r(N - 1)} \tag{3.1}$$

N=number of components under investigation,

r=the average correlation coefficient between the components

Calculation of Cronbach's alpha is quite time-consuming, so it calculates by specialized statistical programs (e.g. SPSS).The system includes a result matrix (usually a result matrix for a specific study): one row indicates one person (one "case"),one column indicates a study query with results for each subject.The software measures the correlation coefficients between the question, then determines the average and measures the Cronbach alpha using the formula. Cronbach's alpha interpretation:

1. 0.9-1 [very good]
2. 0.8-0.9 [good]
3. 0.7-0.8 [sufficient]
4. 0.6-0.7 [questionable]
5. 0.5-0.6 [poor]
6. 0-0.5 [insufficient]

The use of Cronbach's alpha should be with some caution. Let's say someone has developed a test that is not very successful, in which the tasks are poorly correlated with each other.Interpretation – "poor". Most likely, the low average r is due to the fact that the tasks are poorly thought out. We should work out their quality.

3.1.2 KMO-Barlett test

The condition of sphericity for KMO and Bartlett. The Kaiser-Meyer-Olkin (KMO) calculate the selective adequacy used to check the hypothesis that partial

differences between variables are low.Bartlett's sphericity criterion tests the hypothesis that the matrix is an identity matrix[10]. If the hypothesis is true, the model of the factor is inadequate.To order to explain the probability of implementing factor analysis, the Bartlett sphericity test and the Kaiser – Meyer – Olken sample adequacy criteria (KMO statistics) are more commonly used. Bartlett's sphericity criterion provides an opportunity to test the hypothesis that in the general population the variables involved in the analysis are not correlated.The adequacy criterion for the Kaiser – Meyer – Olken sample (KMO statistics) verifies the suitability of the available data for factor analysis. The smaller the value of this criterion, the less likely it is that the relationship between pairs of variables is due to other variables. Large values of the criterion indicate the possibility of explaining the correlation between pairs of variables as a third factor.The use of factor analysis can be considered reasonable.KMO numbers range from 0 to 1, and the closest to cohesion, the more appropriate the usage of factor analysis becomes.Typically, factor analysis is considered appropriate if the KMO statistic exceeds 0.5.The KMO check formula is:

$$MO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} u} \tag{3.2}$$

where:

R=$[r_{ij}]$=correlation matrix;

U=$[u_{ij}]$ is the partial covariance matrix

KMO returns values from 0 to 1. The law of thumb for understanding the figures:

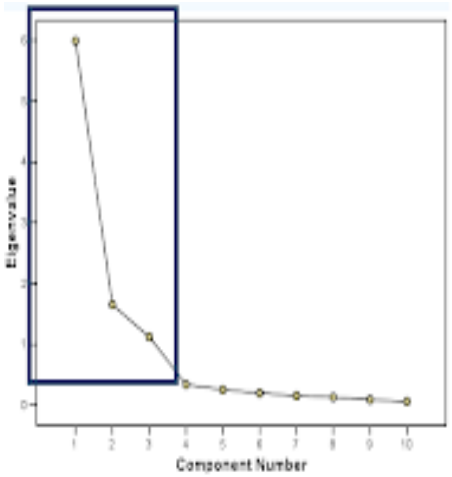
1. KMO values between 0.8 and 1 suggest that the sampling is sufficient..

2. KMO values less than 0.6 indicate that sampling is not appropriate and that remedial action should be taken. Some scholars have set this value to 0.5, but use your own judgement for values between 0.5 and 0.6.

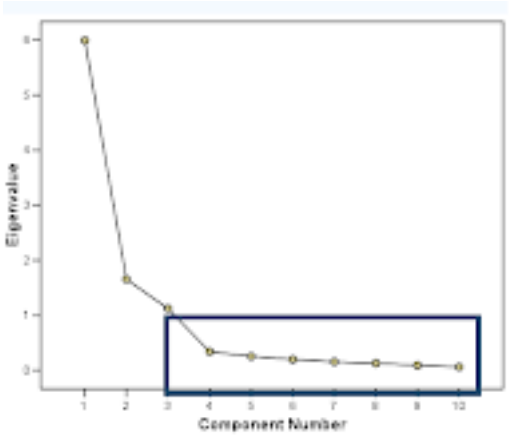
3. KMO Values similar to zero mean that there are broad partial correlations relative to the number of correlations. In other terms, there are widespread that present a big challenge for factor analysis.

3.2 Scree plot

Scree plot is a statistical method used to pick the number of appropriate components or variables to be included in the component analysis or factor analysis.The scree test for the number of factors dates back to Cattell (1966).Cattell stated that the method that contributed to the creation of the scree test was one that contributed him to remove the key components of the correlation matrices and then to look for a "various symbol".A Scree Plot is a basic line section plot that displays the fraction of the overall variation in the results. It is a map, in descending order of magnitude, of the sum of the matrix of the correlation.In the form of a factor analysis or a key variable study, a scree plot allows the observer to imagine the relative value of the variables, a drastic decrease in the plot shows the subsequent variables are overlooked. Programs usually have a default cut-off for the number of generated factors, such as all factors with an eigenvalue of ≥ 1.That is since a factor with its eigenvalue of 1 takes into consideration as many variations as a single variable, and the rationale is that it is worth keeping only certain variables that justify at least the same amount of variations as a single variable.But instead clipping 1 results in more variables than the consumer expected, or removes a potentially significant factor whose eigenvalue is only below 1. Therefore, please use this test with strict care. The Scree Plot has two lines: the lower line displays the variance ratio for each main component, while the upper line shows the total variance shown by the first N components. The main components are sorted in decreasing order of variance, such that the most significant main component is often mentioned first.



The materials on the shallow slope offer no contribution to the solution.



3.3 Variamax rotation

The factor load matrix, sometimes called the factor pattern matrix, is the significant consequence in factor analysis.It includes the coefficients used for the normalization of Transmit variables in terms of influences. These coefficients, known as factor loads, reflect associations between factors and variables.The coefficient with an absolute value implies that the element and the component are strongly related. The factor loading matrix coefficients can be used to describe the variables.

The matrix of the root or non-inverted factors shows the association between the factors and the actual variables, it never refers to factors that can be understood as the factors refer to other variables.When rotating factors, it is ideal that each factor has a non-zero or important load (coefficient) with just a limited number of variables. Similarly, it is optimal for each variable to provide a non-zero or substantial load with a limited number of variables, if possible for one factor. If many variables have large load factor values of the same component, they are difficult to interpret.Rotation shall not impact the generalities or the proportion of overall variation stated.However, the proportion of variation owing to the effect of each element differs.As a result of rotation, the variance explained by each factor was redistributed. Therefore, different rotation methods help interpret different factors.

Rotation is called orthogonal rotation if the rotation preserves a rectangular coordinate system.The most common rotation method is the varimax (dispersion

— maximizing rotation) method.

The varimax condition is the complexity index of each component, which is equal to the number of variables correlated with that element. The "varimax" approach maximizes the distribution of the load squares for each component, resulting in a significant increase and a reduction in the small values of the component loads.As a consequence, a basic structure for each element is obtained separately.

Matrix $A \leftrightarrow_{mxg} v_{ij}$ of the primary factor mapping of the dimension m x g of weight coefficients.Where m is the number of parameters being studied, g-number of common factors.Rotation consists of the following matrix operation:

$$V = A\Lambda$$

$V \leftrightarrow_{mxg} v_{ij}$-skew factor structure;

$\Lambda \leftrightarrow_{gxg} \Lambda_{ij}$-rotation matrix;

Correlation matrix $C \leftrightarrow_{gxg} c_{ij}$ dimensions g x g between final factors when the source data is standardized (variances of variables are equal 1, and the average 0) is calculated using the formula:

$$C = \Lambda^T \Lambda \tag{3.3}$$

For the rotation matrix Λ the following relations must be fulfilled: $c_{ij} = \sum_{i=1}^{g} \lambda_{ij}^2 = 1, (j = 1, ..., g)$.

As well as restrictions on the type of inequality:

$$|c_{ij}| = |\sum_{k=1}^{g} \lambda_{ki}\lambda_{kj}| \leq 1. \tag{3.4}$$

If the resulting factorial solution must be orthogonal, then restrictions of the inequality type, replace with $c_{ij} = \sum_{k=1}^{g} \lambda_{ki}\lambda_{kj} = 0$.

The factor rotation problem corresponds to maximization or minimization of a certain criterion K, as a function of the matrix elements of the resulting factor structure: $K = f(v_{ij}) = f(a_{ij}, \lambda_{ij})$.Since the a_{ij} elements are set, the task reduces to finding the extremum of the function $K = f(\lambda_{ij})$ from the independent variables λ_{ij} with restrictions:

$$\sum_{i=1}^{g} \lambda_{ij}^2 = 1 \text{ and}$$

$$|\sum_{k=1}^{g} \lambda_{ki}\lambda_{kj}| \leq 1 \text{ or } \sum_{k=1}^{g} \lambda_{ki}\lambda_{kj} = 0$$

It is also suggested to use the following restrictions on the type of resulting factor structure. Generalities of the finite factor variables structures must be no larger than the generalities of variables in the original factor structure, and no less than a certain threshold of significance:
$h_i^v = \sqrt{\sum_{k=1}^{g}} \leq \sqrt{\sum_{k=1}^{g} a_{ik}^2} = h_i^a$ and $h_i^v \geq p$.

Varimax formula with math: $K = n\sum_{p=1}^{g}\sum_{i=1}^{m}(\frac{v_{ip}}{h_i})^2 - \sum_{p=1}^{g}(\sum_{i=1}^{m}\frac{v_{ip}^2}{h_i^2})^2$

The maximization of the varimax criterion corresponds to the maximization of variances squares of loads of factors. Thus the theoretical complexity of the factor decreases, factor loads are close to 0 or 1, and the factor can be best how to interpret.Normalization of factor loads in this case criteria eliminates the difference between the contributions of individual parameters in proportion to their generalities.

3.4 Hypothesis testing

Statistical hypotheses are assumptions Researchers about measurement results expressed in formalized concise form.

Testing statistical hypotheses is a process of deciding whether the statistical hypothesis in question contradicts an observable data sample. Often a sample is made to determine the arguments against the hypothesis regarding the population (population)[11]. This process is known as hypothesis testing (statistical hypothesis testing or significance testing); it is a quantitative measure of the arguments against a part.

A statistical test or statistical criterion is a strict mathematical rule by which the statistical hypothesis is accepted or rejected.

5 stages were established when testing hypotheses[11]:

1. Definition of the null (H_0) and alternative hypothesis (H_1) in the study. Determining the level of significance of the criterion.

2. Selection of the necessary data from the selection.

3. Calculation of the statistic value of the criterion corresponding to H_0.

4. Define P-value

5. Interpretation of the achieved level of significance p and results.

Significance level determines the probability of rejecting claim.The percentage value is known as 5%.An significant step in the testing of statistical theories is to assess the degree of statistical significance of the alpha, i.e. the limit permitted by the researcher's likelihood of an incorrect variance of the null hypothesis.
*Critical region:*In order to evaluate whether or not to dismiss the null hypothesis, it is therefore important to establish the crucial region for the estimation of the hypothesis.[11]. *H_O-null hypothesis:*It makes an assumption that differences between compared samples are absent.In fact, differences can deviate from 0 but not be reliable or not proven.
H_A-an alternative hypothesis (opposing the null hypothesis. Its meaning is that the differences between samples there and that they are reliable.Both of null and alternative hypothesis are expressed in terms of a parameter,such as probability or mean.
When checking the significance of a hypothesis, it should be formulated independently of the data used in its verification (prior to the verification). In this case, you can get a really productive result[11].Always test the null hypothesis (H_0), which rejects the effect (for example, the difference in means is zero) in the population. For example, when comparing smoking rates in men and women in a population, the null hypothesis H_0 would mean that smoking rates are the same in women and men in a population. Then an alternative hypothesis (H_1) is determined, which is accepted if the null hypothesis is false[19]. An alternative hypothesis is more relevant to the theory that they are going to research. So, in this example, the alternative hypothesis h1 is to argue that smoking rates are different for women and men in the population.
All statistics of the criterion obey the known theoretical probability distributions. The value of the criterion statistics obtained from the sample is associated with the already known distribution to which it obeys in order to obtain the p value, the area of two "tails" (or one "tail", in the case of a one-tailed hypothesis) of the probability distribution[11]. Most computer packages automatically calculate the two-tailed p.
The value of p is the probability of obtaining our calculated value of the criterion or its even greater value, if the null hypothesis is true.A p is the probability of rejecting the null hypothesis, provided that it is true. The null hypothesis always

refers to a population of greater interest than a sample. As part of a hypothesis test, we either reject the null hypothesis and accept the alternative, or do not reject the null hypothesis[11].

Application of p value

1.It is traditionally believed that if p<0.05, ($\alpha = 0.05$), then the arguments are enough to reject the null hypothesis, although there is little chance against this. Then we can reject the null hypothesis and say that the results are significant at the 5% level.

2.On the contrary, if p>0.05, then the arguments are not enough to reject the null hypothesis. Without rejecting the null hypothesis, it can be stated that the results are not significant at the 5% level. This conclusion does not mean that the null hypothesis is true, there are simply not enough arguments (perhaps a small sample size) to reject it[11].

One-tailed testA statistical criterion that takes into account a priori knowledge about the direction (increase or decrease) of the value of the studied parameter of one group in relation to the same parameter of another group.

Two tailed test A hypothesis test condition determined prior to data collection.A two-way test does not imply that the direction of the shift in the value of the analyzed parameter of one group relative to another is known in advance.A two-way test is more universal than a one-way test. It is also more conservative compared to a one-way test.The minimum difference between groups required for a one-way test is less than for a two-way test.

3.4.1 T-test

The t-test was developed by William Gosset (1876-1937) to assess the quality of beer in Guinness breweries in Dublin (Ireland). In connection with the obligations to the company for the non-disclosure of trade secrets (Guinness management considered the use of the statistical apparatus in their work), Gosset's article was published in 1908 in the journal "Biometrics" under the pseudonym "Student" (Student).William Gosset is an English mathematician and chemist who, after graduation University began working at the Guinness factory, engaged in quality control in the process of creation beer. Why was a nickname Student chosen? One version says that the alias student was used for Guinness, because the company itself wanted keep secret statistics working on it. Student's criterion

(Student's t-test or just "t-test") is used if you want to compare only two groups of quantitative attributes with a normal distribution (a special case of analysis of variance).Note: this criterion cannot be used when comparing several groups in pairs, in this case, analysis of variance is necessary. Erroneous use of Student's criterion increases the likelihood of "revealing" non-existent differences.For example, instead of recognizing several treatments as equally effective (or ineffective), one of them is declared the best.

Two events are called independent if the advance of one of them does not affect the advance of the other. Similarly, two sets can be called independent if the properties of one of them are in no way connected with the properties of the other.

Student's t-test - the general name for statistical tests in which the statistic of a criterion has a student distribution. Most often, t-criteria are used to verify the equality of mean values in two samples.The null hypothesis assumes that the averages are equal (the negation of this assumption is called the shift hypothesis). Student t-test for independent samples is used to compare the average values of two independent samples. Terms of use:

1. The compared values do not constitute a pair of correlating values.
2. The distribution of characteristics in each sample corresponds to the normal distribution.
3. Characteristic variances in the samples are approximately equal (checked using the F-Fisher criterion).

Student's t criterion is aimed at assessing differences in the values of the average $\overline{X}$ and $\overline{Y}$ has two samples X and Y, which are distributed according to the normal law. One of the main advantages of the criterion is the breadth of its application. It can be used to compare the averages of connected and disconnected samples, and the samples may not be equal in magnitude.The basic formula for calculating statistics for testing the hypothesis of comparing two independent samples has the following view. If the sample size is small (n <30), then the statistics have a Student's distribution. In the general case, the formula for calculating by t - Student's criterion is as follows:

$$t_e = \left| \frac{\overline{X} - \overline{Y}}{Sd} \right|$$

где $Sd = \sqrt{S_x^2 + S_y^2}$.First, consider the equal samples. In this case, n1 = n2 = n.

$$Sd = \sqrt{S_x^2 + S_y^2} = \sqrt{\frac{\sum (x_i - \overline{x})^2 + \sum (y_i - \overline{y})^2}{(N-1) \times n}} \tag{3.5}$$

In the case of unequal samples of $n_1 \neq n_2$, the expression will be calculated as follows:

$$Sd = \sqrt{S_x^2 + S_y^2} = \sqrt{\frac{\sum (x_i - \overline{x})^2 + \sum (y_i - \overline{y})^2}{n_1 + n_2 - 2} * \frac{n_1 + n_2}{n_1 * n_2}} \tag{3.6}$$

In both cases, the calculation of the number of degrees of freedom is carried out according to the formula:

k= $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ where n1 and n2 are the values of the first and second samples, respectively. It is clear that with numerically equal samples, k = 2 ×(n - 2).One of the main advantages of the criterion is the breadth of its application. Use of this The method is common in many areas of medicine, bioengineering, biochemistry, agriculture, retail, law, etc. For example, a white blood cell count is measured in healthy animals, and then in the same animals after exposure to a certain dose of radiation.

3.5 Python

A programming may be described as a series of sequential commands (algorithms) for an entity (performer) that must execute them in order to accomplish a certain objective.And you can "program" a individual, if you give him a "how to cook pancakes" tutorial, and he begins to do it easily.The Python programming language was created by Dutch Guido van Rossum in 1991.Its name – Python (or Python) - got from the name of the TV series, not the reptile.Python is still under active development. New releases are periodically published. There are two versions supported:Python 2.x and Python 3.x. This is the English letter "x" that signifies a particular update.There's a small gap between the second and the third Python.Python is an interpreted programming language.This means that the source code is partially converted to machine code when it is read by a special interpreter program.Python is a fully-fledged, nearly standardized programming language employed in a number of fields. The key, though not the only, method-

ology that it embraces is object-oriented programming. In this course, however, we will only consider artifacts and review structural programming, as it is the foundation.

Mathematical models of principal component analysis and factor analysis. Interpretation of factors. Example of factor analysis in Python. Factor loads, factor labels, and their interpretation. Rotation of factors.Python allows you to solve the problem of automating data collection and processing, speeds up data analysis, and allows you to implement new approaches to analysis, such as solving problems using neural network training.

4. Main parts

4.1 Data processing

Primary data is collected using questionnaire and provided qualitative data.The main purpose of this study is an attempt to reduce the number of factors and define main aspects of the resulted construct.The survey is prepared by using various sources like [26], [25],[9] and adapted to the context of Kazakhstan.An online survey questionnaire is consisting 27 questions. The survey is prepared languages,namely Kazakh, Russian, and English and send out to students from 16 universities within the country and received 314 students participants.Students took as a sample through non probability convenience sampling technique.First part of questionnaire is directed to collect demographic data and items related to major selection were in the second part of the survey.Table 1 provides demographic information on participants.The number of respondents in the Kazakh language is 109,in Russian 93 and in English 112.

Table 1

Language	Age group	Gender	University GPA
Kazakh (34.7%)	16-18y (18.5%)	Male (45.9%)	3.5-4.0 (43.3%)
Russian (29.6%)	19-21y (42.7%)	Female (54.1%)	2.5-3.4 (46.2%)
English (35.7%)	22-24y (25.1%)		1.5-2.4 (9.55)%
	24-more (13.7%)		1.0-1.4 (0,95%)

Most respondents are between 19-21 years old students. They took 42.7% (134 students) from total. However, 24 years or older than 24 years are 13.7% from 314 students. Students of the engineering speciality took part in the survey by 21.3% (67students) and it is the highest frequency of respondents. Then comes students majoring in pedagogy and mathematics with 20% (63 students). The minimum size of the surveyed participants are attended by journalists and fine and applied art. They took only 1% (3 students from each).

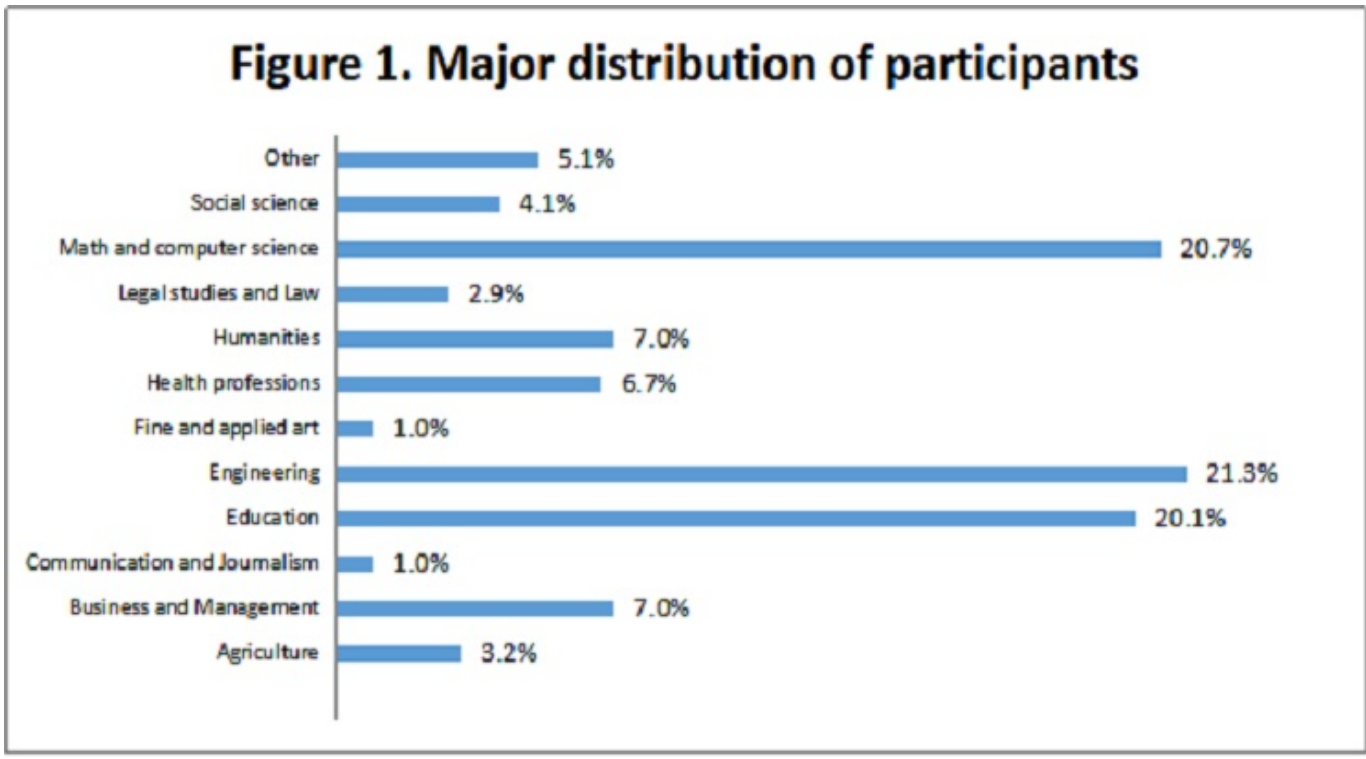
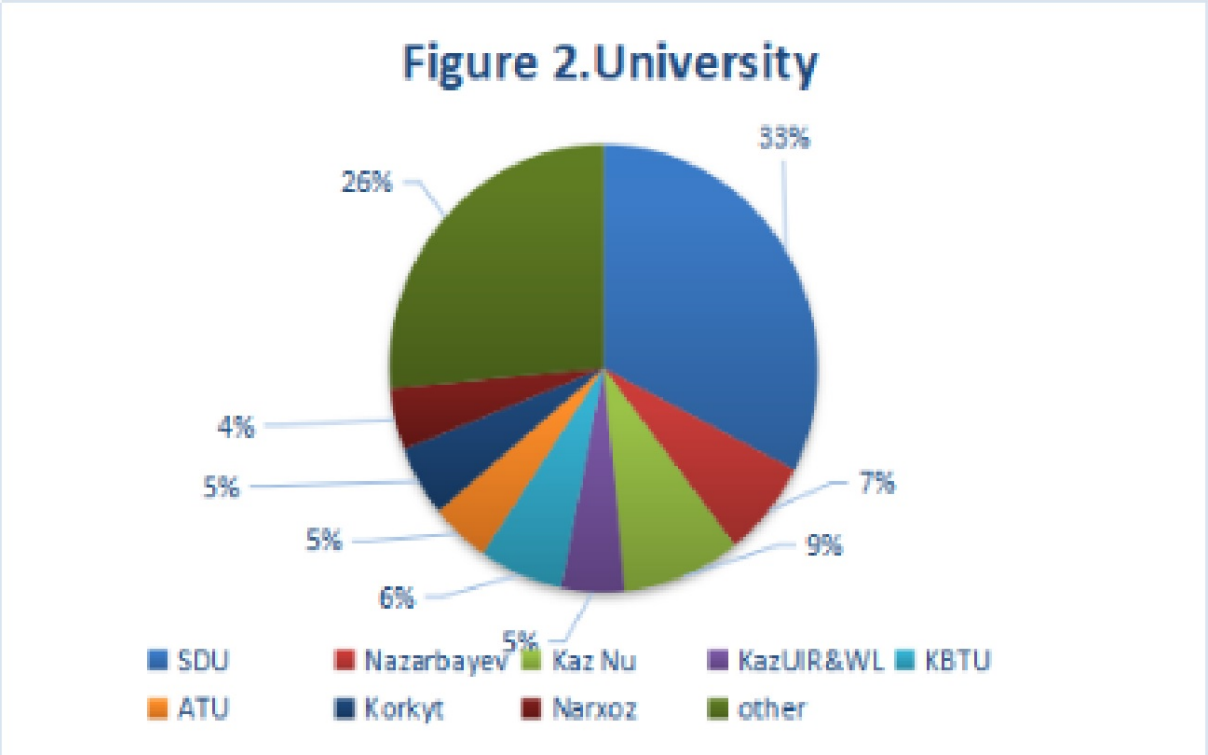


Figure 2 provides information about the number of students surveyed by more than 16 universities. The majority of respondents are students of SDU university. Universities with a less number of participants grouped together,their number is 83 and named as an other in the Figure 2.



4.1.1 Instruments

All quantitative data were first collected in excel format, then the answers of the University, GPA, Gender, Age group were replaced by numbers and stored in a separate file to prepare for analysis.New quantitative data has been created. Participants answered to questions by online form. Responses are evaluated using the Likert Scale. Items are graded from 1 to 5 points. Accordingly, 1-'strongly disagree', 2-'disagree', 3-'neutral',4-'agree',5-'strongly disagree'. The answers are translated into Kazakh and Russian languages accordingly with this grading system. In order to check the internal consistency of scale items, Chronbach alpha reliability analysis is performed.For dimensionality reduction factor analysis through Principal Component Analysis is implemented with Varimax rotation. The Kaiser-Meyer Olkin sampling Adequacy index is a figure showing the proportion of variation in your variables that could be caused by underlying factors. Scree plot is used to map the data's eigen values and to calculate the amount of variables of the key components. When utilizing rotation techniques such as VARIMAX, we have additional resources that make it simpler to analyze the variables and will also increase the validity of the tests.T-Test are applied to get answers for study questions by testing a number of hypotheses between different elements of demographic data and main factors on selection of study field.T-test is

applied when we measures difference between means of two group.We compared factors for gender differences.We used t-test to compare men's versus women's means to determine level of significance.Also,we tested main factor differences according to SDU students and other local universities.

4.2 Results

In applying the Kaiser-Meyer-Olkin's (KMO) overall measure of sampling adequacy (MSA), a score of 0.850 is recorded which is in the acceptable range based on a KMO overall MSA greater than 0.60 being considered acceptable, [29].Kaiser-Meyer-Olkin (KMO) Barlett's test of sphericity threshold is high and a high significant chi-square (χ^2= 2102.9 (1 d. p.), p<0.0001).Chronbach's Alpha reliability was performed ,to check consistency of the scale items.Particular sample with the value of 0.856 Chronbach's Alpha shows a high level of internal consistency for our scale.

PCA finds principal components in descending order of variations explained. The first components account for more variations than the later ones. The 1st principal component accounts for the maximum amount of variations possible in data, and the 2nd principal component extracts the maximum possible variations in data after excluding what was explained by the 1st component. Extractions can be done until all the by the last principal variations are accounted for components.Covariance matrix allows for dataset variation, and dataset variation is a way to quantify how much knowledge we have in the data.Next tables show us calculation values of our covariance matrix.

1. University costs played a major role in my choice	1.000.000	0.139450	0.077363	0.063835	0.098701	0.056535	0.054784	0.081677	0.085414
2. I was influenced by my parents in my choice	0.139450	1.000.000	0.229133	0.117678	0.152878	0.123594	0.071863	0.243547	0.150725
3. My high school career advisor influenced my choice	0.077363	0.229133	1.000.000	0.086126	0.343229	0.333795	0.138461	0.024553	0.254188
4. The job's accessibility affected my choice	0.063835	0.117678	0.086126	1.000.000	0.092913	0.156192	0.106823	0.282124	0.247282
5. My religious convictions influenced the selection of my choice	0.098701	0.152878	0.343229	0.092913	1.000.000	0.235325	0.117624	0.045280	0.154693
6. I was encouraged by a teacher because I was good at my main subjects.	0.056535	0.123594	0.333795	0.156192	0.235325	1.000.000	0.130232	0.030406	0.260689
7. My own life experience affected me in my choice (eg. I want to be a doctor, because a doctor saved someone's life in my family)	0.054784	0.071863	0.138461	0.106823	0.117624	0.130232	1.000.000	0.202533	0.240103
8. Upon graduation good salary affected my choice	0.081677	0.243547	0.024553	0.282124	0.045280	0.030406	0.202533	1.000.000	0.288551
9. My academic performance at High School affected my choice	0.085414	0.150725	0.254188	0.247282	0.154693	0.260689	0.240103	0.288551	1.000.000
10. Duration of study influenced my choice (e.g. the major will require further training like a master's degree).	0.093534	0.204558	0.190320	0.233251	0.100048	0.258876	0.315225	0.258717	0.312857
...	...	...	...	...	...	...	...	...	...
18. Presentations of currently enrolled students made a great impact in my choice	0.139979	0.140363	0.252728	0.224492	0.189478	0.236278	0.300645	0.263400	0.326597
19. Alumni's presentations influenced my choice	0.091478	0.143827	0.217700	0.151761	0.199071	0.231127	0.279040	0.309775	0.240697
20. Prestige of profession affected my selection	0.082216	0.051319	-0.021555	0.258971	0.090090	0.059978	0.193719	0.420661	0.260274
21. I believe that professionals in this field can help develop my country	0.005259	-0.037625	0.033004	0.207143	-0.034891	0.022309	0.079470	0.098686	0.081552
22. My UNT result was important when I selected my major	0.031632	0.047021	0.123637	0.129088	0.166916	0.188283	0.026754	0.101027	0.170727
23. My high school teacher asked me to specialize in this field as it has high demands nowadays	0.079384	0.172535	0.247915	0.007810	0.306244	0.369279	0.192714	0.126482	0.215373
24. Opinions of my peers affected my selection	0.097239	0.276551	0.286231	0.035389	0.275248	0.223841	0.198853	0.147670	0.223152
25. Current situation in my family affected my selection	0.113007	0.288276	0.262195	0.020298	0.248068	0.265128	0.097540	0.159319	0.225607
26. Famous personalities who had the same specialization in that field affected my major selection	0.051955	0.055098	0.174735	0.233380	0.286580	0.265855	0.224331	0.188149	0.213341
27. My skills were major effect in my choice	-0.005809	-0.062067	-0.015267	0.130855	0.044467	0.191910	0.187544	0.104140	0.187633

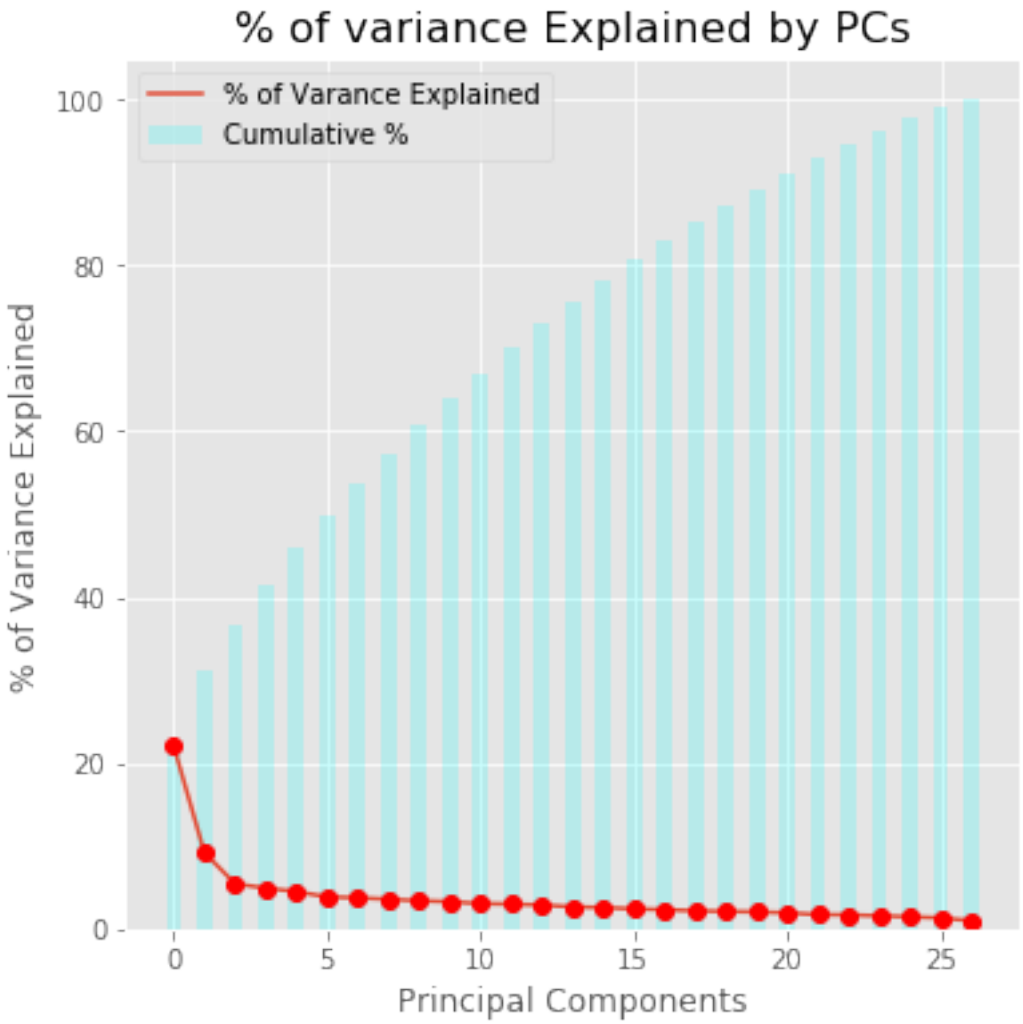
0.093534	0.016383	0.109156	0.068387	0.036120	0.118966	0.016107	0.029684	0.139979	0.091478	0.082216
0.204558	0.429881	0.187837	-0.072523	0.323714	0.092451	-0.215645	0.105721	0.140363	0.143827	0.051319
0.190320	0.215423	0.177067	-0.005348	0.199713	0.135434	-0.090876	0.175086	0.252728	0.217700	-0.021555
0.233251	0.034554	0.070036	0.183430	0.030796	0.137033	0.131487	0.176577	0.224492	0.151761	0.258971
0.100048	0.161984	0.104400	-0.036931	0.315963	0.113190	-0.071626	0.196791	0.189478	0.199071	0.090090
0.258876	0.231827	0.161600	0.063116	0.232761	0.073267	0.043126	0.124302	0.236278	0.231127	0.059978
0.315225	0.159050	0.172150	0.153136	0.210291	0.089675	0.168781	0.260886	0.300645	0.279040	0.193719
0.258717	0.170574	0.164297	0.198299	0.232858	0.135847	0.070852	0.206976	0.263400	0.309775	0.420661
0.312857	0.163535	0.084581	0.238810	0.231727	0.184232	0.108422	0.166417	0.326597	0.240697	0.260274
1.000.000	0.270335	0.269603	0.175994	0.333001	0.245151	0.051531	0.239771	0.320553	0.309641	0.267793
...	...	...	...	...	...	...	...	...	...	...
0.320553	0.399107	0.285434	0.205194	0.352583	0.275433	0.099004	0.493877	1.000.000	0.650251	0.386430
0.309641	0.373133	0.299285	0.138730	0.392747	0.309336	0.128242	0.394458	0.650251	1.000.000	0.305249
0.267793	0.149984	0.168916	0.281082	0.189398	0.160288	0.174468	0.282126	0.386430	0.305249	1.000.000
0.088004	-0.069419	0.002854	0.227551	-0.033821	0.035327	0.268571	0.146030	0.041645	0.145934	0.212465
0.111522	0.108125	0.157676	0.068994	0.128252	0.320912	-0.095817	0.117196	0.183072	0.184686	0.145145
0.243282	0.337287	0.134029	0.047694	0.332244	0.225622	-0.062166	0.246188	0.319575	0.309319	0.161515
0.228575	0.331799	0.288230	-0.004622	0.388173	0.210622	0.042699	0.313962	0.393421	0.427515	0.148694
0.208842	0.370163	0.364189	-0.038957	0.327434	0.220662	-0.183779	0.174013	0.273888	0.266954	0.130710
0.249990	0.125345	0.163059	0.113019	0.331229	0.223865	0.179563	0.236718	0.370869	0.325222	0.304829
0.164664	0.046604	0.007321	0.244470	0.083067	0.033817	0.422534	0.216020	0.262420	0.190773	0.246538

0.005259	0.031632	0.079384	0.097239	0.113007	0.051955	-0.005809
-0.037625	0.047021	0.172535	0.276551	0.288276	0.055098	-0.062067
0.033004	0.123637	0.247915	0.286231	0.262195	0.174735	-0.015267
0.207143	0.129088	0.007810	0.035389	0.020298	0.233380	0.130855
-0.034891	0.166916	0.306244	0.275248	0.248068	0.286580	0.044467
0.022309	0.188283	0.369279	0.223841	0.265128	0.265855	0.191910
0.079470	0.026754	0.192714	0.198853	0.097540	0.224331	0.187544
0.098686	0.101027	0.126482	0.147670	0.159319	0.188149	0.104140
0.081552	0.170727	0.215373	0.223152	0.225607	0.213341	0.187633
0.088004	0.111522	0.243282	0.228575	0.208842	0.249990	0.164664
...	...	...	...	...	...	...
0.041645	0.183072	0.319575	0.393421	0.273888	0.370869	0.262420
0.145934	0.184686	0.309319	0.427515	0.266954	0.325222	0.190773
0.212465	0.145145	0.161515	0.148694	0.130710	0.304829	0.246538
1.000.000	0.048390	-0.014758	-0.016473	-0.041113	0.121245	0.149971
0.048390	1.000.000	0.223776	0.173081	0.178317	0.221747	-0.022294
-0.014758	0.223776	1.000.000	0.304988	0.293784	0.350680	0.167863
-0.016473	0.173081	0.304988	1.000.000	0.461545	0.271696	0.151708
-0.041113	0.178317	0.293784	0.461545	1.000.000	0.164164	0.015684
0.121245	0.221747	0.350680	0.271696	0.164164	1.000.000	0.147877
0.149971	-0.022294	0.167863	0.151708	0.015684	0.147877	1.000.000

So, we decided to consider ([13]) the first 6 factors which resulted in 27 Items.From the next table Total Variance Explained it is clear that the 49,97% of the variance is explained by the first six components.

PC		Eigenvalue	% of Variance Exp	Cumulative %
0	0	5.951	22%	22%
1	1	2.485	9%	31%
2	2	1.472	5%	37%
3	3	1.327	5%	42%
4	4	1.217	5%	46%
5	5	1.041	4%	50%
6	6	1.014	4%	54%
7	7	0.967	4%	57%
8	8	0.919	3%	61%
9	9	0.885	3%	64%
10	10	0.825	3%	67%
11	11	0.809	3%	70%
12	12	0.779	3%	73%
13	13	0.718	3%	76%
14	14	0.696	3%	78%
15	15	0.656	2%	81%
16	16	0.624	2%	83%
17	17	0.591	2%	85%
18	18	0.570	2%	87%
19	19	0.556	2%	89%
20	20	0.524	2%	91%
21	21	0.467	2%	93%
22	22	0.458	2%	95%
23	23	0.427	2%	96%
24	24	0.397	1%	98%
25	25	0.358	1%	99%
26	26	0.267	0.990%	100%

A scree plot and a bar chart for the cumulated percentage of variance are drawn in the same graph as shown on Figure.



The next analysis shows how factor loadings. Among the principal components (PCs),at the beginning only the first 6 are selected.The loadings matrix in output shows the relationship between old variables with new principal components by calculating the coordinate of the old variables along the PC (principal component) axes:

	PC1	PC2	PC3	PC4	PC5	PC6
0	0.182	-0.025	-0.127	0.184	0.137	0.530
1	0.376	-0.399	-0.330	-0.049	0.445	-0.084
2	0.426	-0.295	0.314	0.195	0.256	0.143
3	0.319	0.360	-0.148	0.396	0.291	-0.200
4	0.413	-0.259	0.382	0.175	0.059	-0.241
5	0.458	-0.108	0.459	0.173	0.243	0.138
6	0.432	0.214	0.059	-0.253	0.134	0.071
7	0.442	0.244	-0.471	0.029	0.216	-0.287
8	0.505	0.173	0.029	0.226	0.341	0.072
9	0.561	0.101	-0.132	0.011	0.218	0.118
10	0.553	-0.341	-0.197	-0.248	0.042	0.011
11	0.475	-0.176	-0.270	-0.092	-0.169	0.297
12	0.254	0.497	-0.161	0.150	-0.028	0.308
13	0.609	-0.253	-0.028	-0.220	0.024	-0.213
14	0.434	-0.034	-0.187	0.351	-0.429	0.214
15	0.070	0.660	0.308	-0.216	0.052	0.123
16	0.548	0.138	-0.014	-0.209	-0.223	-0.126
17	0.726	0.130	-0.019	-0.173	-0.196	0.022
18	0.695	0.088	-0.032	-0.197	-0.246	-0.005
19	0.478	0.432	-0.262	0.027	-0.061	-0.227
20	0.134	0.480	0.009	0.153	0.040	0.004
21	0.339	-0.086	0.020	0.561	-0.388	-0.079
22	0.553	-0.189	0.314	0.042	-0.080	-0.084
23	0.606	-0.228	0.117	-0.215	-0.089	0.077
24	0.525	-0.408	-0.060	-0.004	0.004	0.156
25	0.542	0.156	0.267	0.163	-0.137	-0.297
26	0.295	0.492	0.318	-0.289	0.054	0.138

PCA often needs rotation for easier interpretation. The current we used the most popular method called Varimax rotation. Varimax orthogonal rotation tries to increase the variation of square loads in each factor such that each factor has just a few variables with high loads and several other variables with small loads [26], Only loadings greater than |0.40| are considered. Results of Varimax rotation is shown on Table 4. From rotated component matrix, we eliminated questions 7,10,13,21,26 (Appendix 1) with lowest loadings. We obtained components with a Chronbach's alpha greater than 0,6.Components 1,2,3 are satisfied to condition. We considered each elements in components 4,5,6 separately, because these components reliability scale less than 0,6.

Factor 1:External Influences External influence factor has 8 items .As a result, external factors play an important role in choosing a profession for a child. It has loadings from 0.682 to 0.489. All these factors are more related to external influences like influence of peers and relatives, and situation of family, also advertisement of specializations. Reliability scale is 0,810(Chronbach alpha).
Factor 2:Teacher influences Second component gave information that students can influence by school teachers. Reliability scale is 0,638.It is given with loadings 0.692 to 0.509. That is why,parents should make sure that the teacher is a person of good level. Always be in close contact with the teacher.It has 4 items
Factor 3:Influence of occupation salary The third is important component and it covers job accessibility and prestige of major, also salary. Also, it has 4 items. The child thinks that studying for a prestigious and popular specialty will be received on the highest salary. Influence of occupation salary factor is given with loadings ranging from 0.702 to 0,565.Chronbach's alpha is 0,624. Since the Chronbach alpha of the other 3 components is very low,we considered each element as a separate factor.Components between 4 and 9 covers only one factor of the study that is:
Factor 4:Personal interest influences
Factor 5.: Personal skill influences
Factor 6: Parent's affect
Factor 7:National test affect
Factor 8.State grant affect factor
Factor 9.University cost affect factor

Table 4. Rotated Component Matrix

	Component					
	1	2	3	4	5	6
19.Alumni's presentations influenced my choice	0,682					
18.Presentations of currently enrolled students made a great impact in my choice	0,670					
11.My relatives affected my choice	0,647					
14.I wanted to follow my parents' footsteps	0,628					
24.Opinions of my peers affected my selection	0,617					
17.I was influenced by various advertisement sources (e.g. news, social media, etc)	0,559					
12.University location can be considered as a factor which affected my choice	0,538					
25.Current situation in my family affected my selection	0,489					
6.I was encouraged by a teacher because I was good at my main subjects.		0,692				
3.My high school career advisor influenced my choice		0,647				
5.My religious convictions influenced the selection of my choice		0,612				
23.My high school teacher asked me to specialize in this field as it has high demands nowadays		0,509				
8.Upon graduation good salary affected my choice			0,702			
4.The job's accessibility affected my choice			0,689			
20.Prestige of profession affected my selection			0,565			
9.My academic performance at High School affected my choice			0,460			
16.My personal interest was the strongest factor when choosing a major				0,758		
27.My skills were major effect in my choice				0,675		
2.I was influenced by my parents in my choice				-0,481		
22.My UNT result was important when I selected my major					0,707	
15.The major which I had selected provided more state grants than others					0,635	
1.University costs played a major role in my choice						0,608

As we mentioned before the one of the aim was to determine the most influential factors for Kazakhstani students on the major selection.To answer this question,we compared the arithmetic mean of the 9 main factors which we found in the previous study and selected those more than 3 as our items scaled between 1-5.The 6 factors were found the most important to the Kazakhstani students and the main factors significance was found as acceptable,eventually the results are illustrated in Table.

[Descriptive statistics]

	N	Minimum	Maximum	Mean
External impact	314	1,00	5,00	2,6656
Teacher impact	314	1,00	5,00	2,7428
Parent impact	314	1,00	5,00	2,9968
State grant impact	314	1,00	5,00	3,1688
Cost of the university impact	314	1,00	5,00	3,2452
Occupation salary impact	314	1,00	5,00	3,3599
The national test result impact	314	1,00	5,00	3,3694
Personal skills impact	314	1,00	5,00	3,6975
Personal interest impact	314	1,00	5,00	3,7803

If p-value ≤ 0.05,null hypotheses rejected, in this case difference is highly significant.Otherwise,it was acceptable.T-test results for gender differences table presents hypothesis testing results between gender and main factors.

Hypothesis	Grouping variable	Test applied	p^*
$H_0 1$:There is no differences between impact of national test result factor among males and females	Gender	T-test	0,572
$H_0 2$:There is no differences between the impact of cost of the university factor on major selection between males and females	Gender	T-test	0,156
$H_0 3$:There is no differences between the impact of occupation salary factor on major selection between males and females	Gender	T-test	0,767
$H_0 4$:There is no differences between the impact of personal interest on major selection between males and females	Gender	T-test	0,204
$H_0 5$:There is no differences between the impact of state grant factor on major selection between males and females	Gender	T-test	0,616
$H_0 6$:There is no differences between the impact of personal skills on major selection between males and females	Gender	T-test	0,174

Hypotheses 1,2,3,4,5,6: As p-value$\geq$0.05,so it is statistically insignificant.Thus hypotheses is accepted which means influence of all this 6 factors are not different among males and females.We also decided to see whether there were any differences that affect the selection of Suleyman Demirel University and respondents from other universities which we grouped together.Results are shown on the next Table.All testing results reveal that the mean score between the university groups is not significantly different.

[T-test results for universities]

Hypothesis	Grouping variable	Test applied	p^*	Result
$H_0 1$:There is no differences between the SDU students and other university students by the impact of university cost factor	University	T-test	0,941	Failed to reject
$H_0 2$:There is no differences between the SDU students and other university students by the impact of state grant factor	University	T-test	0,723	Failed to reject
$H_0 3$:There is no differences between the SDU students and other university students by the impact of occupation salary	University	T-test	0,379	Failed to reject
$H_0 4$:There is no differences between the SDU students and other university students by the impact of personal interest	University	T-test	0,157	Failed to reject
$H_0 5$:There is no differences between the SDU students and other university students by the impact of personal skills	University	T-test	0,101	Failed to reject
$H_0 6$:There is no differences between the SDU students and other university students by the impact of national test result factor	University	T-test	0,908	Failed to reject

5. Discussion and Limitation

Factors observed in this research give identical findings to [25], given the fact that two researchers used samples from separate test fields and countries (Kazakhstan, Pakistan). The similarities include the factors such as personal interest affect, skills affect, occupation salary affect, teacher affect, external affect, parent's affect. Nonetheless, owing to the fact that the education structure of the two countries is entirely different, [25] did not find any of the evidence as follows UNT results, state grant, cost of the university.W. Krampton is another analysis that will confirm the efficacy of our research. We did not function with a reduction in dimensionality, but we identified important factors that affect the decision.With respect to the option of majors taken by graduates, our results are broadly compatible with those of earlier study[33].It was noticed that promised jobs, personal value and anticipated earnings after graduation were the most important element in Turkey's main choice of students. Findings of these indicated that, both the preferences of male and female students were affected by variables very close to our findings[33].The finding of our research is that men and women tend to be more attuned to influences that affect their main decision which is same results with research such as [33].In addition, as a result of our research, the majority of Suleyman Demirel's students participated in the survey, we checked the influence of factors on students of SDU and other universities, and found that there was no difference between them.Therefore, the factors affect students in Kazakhstan in approximately the same way. One shortcoming of the study is that majority of participants were from ony city, Almaty. However, we note that Almaty is the largest city in Kazakhstan with many major universities.To order to boost the generalizability of the analysis, it will be repeated in other universities from various areas. The biggest disadvantage of the our survey is that one third of participants are SDU students. That can change a lot of results. Also, reliability

scale of component 2,3 are somewhat low.The results may be improved by increasing number of students participating in the survey and ensuring that there are different universities and majors . The manner in which the data were obtained also restricted the analysis. Subjects were required to enrol to take part in the analysis and to do it online at their own convenience.Administering this sort of survey, if conducted in person, might be more effective.If any doubt exists as to the effect suggested, making a researcher available to address questions or explain the factor mentioned might provide more reliable data which would contribute to more reliable findings. Academic consideration factor was not provided in our work. It includes course description, instructors.The way the data were gathered also restricted the analysis.Subjects were required to apply to engage in the analysis and to do it online for their own convenience. Administering this sort of survey may have been more effective if conducted in person..Furthermore,most of the participants in the study are from universities that allocate a lot of grants to certain specialties.It is better to include more paid students in the study on the suspicion that there were fewer paid students.

6. Conclusion

The study gave construction of a multilingual questionnaire.We applied this multilingual construction to identify main factors to analyse data using Python programming languages.Also, this study shows all the steps needed for PCA along with Python code beyond varimax rotation and presents other tools with descriptions.We determined the relationship between main factors and demographic data. Therefore, this would help anyone who wants to run PCA at a deeper level.Teachers and parents will use the findings of this research to concentrate their attention on helping students making a big decision.The five-scale translated questionnaires are proved in the Appendix.More work on gender gaps on the variables in the choice of main may be helpful.To sum up, researchers should follow up on students who are now operating to see how their standards about their main option have been fulfilled and whether their demographic data should affect their decision.

A. Appendix A

	English	Russian	Kazakh
1	University costs played a major role	Плата за обучение в университете сыграла большую роль в моем выборе	Менің таңдауымда университеттің оқу ақысы үлкен рөл атқарды
2	I was influenced by my parents	На меня повлияли мои родители в моем выборе	Менің таңдауыма ата-анам әсер етті
3	My high school career adv isor influenced my choice	Консультант по карьере в моей школе повлиял на мой выбор	Менің таңдауыма мектебімдегі мамандық таңдау бойынша кеңесші әсер етті
4	The job's accessibility affe cted my choice	Доступность работы повлияла на мой выбор	Жұмыстың қолжетімділігі таңдауыма әсер етті
5	My religious convictions i nfluenced the selection of my major	Мои религиозные убеждения повлияли на мой выбор	Менің діни сенімдерім таңдауыма әсер етті
6	I was encouraged by a teacher because I was good at my main subjects.	На меня повлиял учитель, потому что я был хорош в своих основных предметах.	Маған мектеп мұғалімі әсер етті, себебі мен негізгі пәндерден жақсы болдым.
7	My Life Experiences have affected me (eg. You want to be a doctor, because a doctor saved someone’s	Мой собственный жизненный опыт повлиял на мой выбор (напр. Я	Менің өмірлік тәжірибем таңдауыма әсер етті (мысалы,. Мен дәрігер болғым келеді,

	life in your family)	хочу быть врачом, потому что врач спас чью-то жизнь в моей семье)	өйткені дәрігер менің отбасымдағы біреудің өмірін сақтап қалды)
8	Upon graduation good sal ary affected my choice	Наличие хорошей зарплаты после окончания учебы повлияло на мой выбор	Оқуды аяқтағаннан кейін жақсы жалақы алу мүмкіндігі таңдауыма әсер етті
9	My academic performance in High School affected the selection	Моя успеваемость в средней школе повлияла на мой выбор	Менің орта мектептегі үлгерімім таңдауыма әсер етті
10	Duration of schooling (e.g . the major will require fur ther training like a master' s degree)	Продолжительност ь обучения повлияла на мой выбор (например, профессия потребует дальнейшего обучения, как степень магистра).	Оқу ұзақтығы таңдауыма әсер етті (мысалы, мамандық магистр дәрежесі секілді одан әрі оқуды талап етеді
11	Extended family members affected my selection	Мои родственники повлияли на мой выбор	Менің туыстарым таңдауыма әсер етті
12	University location can be considered as a factor which affected my selection	Расположение университета можно рассматривать как фактор, повлиявший на мой	Университеттің орналасуын таңдауыма әсер еткен фактор ретінде қарастыруға болады

		выбор	
13	Reputation of the university was important for me	Репутация университета была важна для меня в моем выборе	Таңдауым үшін университеттің беделі маңызды болды
14	I wanted to follow my parents footsteps	Я хотела пойти по стопам родителей	Мен ата- анамның ізімен жүргім келді
15	The major which I had selected provided more state grants than others	Профессия, которую я выбрал, давала больше государственных грантов, чем другие	Басқаларға қарағанда мен таңдаған мамандық бойынша көбірек мемлекеттік гранттар берілді
16	My personal interest was the strongest factor when choosing a major	Мой личный интерес был самым сильным фактором при выборе специальности	Менің жеке қызығушылығым мамандықты таңдауда ең үлкен фактор болды
17	Academic assessment of the major that I had selected based from printed or online information	На меня повлияли различные источники рекламы (например, новости, социальные сети и т. д)	Маған түрлі жарнама көздері әсер етті (мысалы, жаңалықтар, әлеуметтік желілер және т. б.)
18	Presentations of currently enrolled students made a great impact	Презентации зачисленных студентов оказали большое влияние на мой выбор	Қазіргі таңда сол мамандық бойынша оқып жатқан студенттерінің презентациялары

			таңдауыма үлкен әсер етті.
19	Alumni's presentations influenced my choice	Презентации выпускников повлияли на мой выбор	Түлектердің презентациялары таңдауыма әсер етті
20	Prestige of profession affected my selection	Престиж профессии повлиял на мой выбор	Мамандықтың беделі таңдауыма әсер етті
21	I assumed that professionals in this field can help develop my country	Я считаю, что профессионалы в этой области могут помочь развитию моей страны	Менің ойымша, осы саладағы мамандар еліміздің дамуына көмектесе алады
22	My school graduation exam result was important when I selected my major	Мой результат ЕНТ был важен, когда я выбирал свою специальность	Мамандығымды таңдағанда ҰБТ-ның нәтижесі маңызды болды
23	My high school teacher asked me to specialize in this field as it has high demands nowadays	Мой учитель средней школы попросил меня специализироватьс я в этой области, поскольку она имеет высокие требования в настоящее время	Менің орта мектеп мұғалімім маған осы саланы меңгеруге кеңес берді, себебі ол қазіргі уақытта жоғары сұраныста бар сала
24	Opinion of my peers affected my selection	Мнения моих сверстников	Менің құрдастарымның пікірлері таңдауыма

		повлияли на мой выбор	әсер етті
25	Current situation in my family affected my selection	Текущая ситуация в моей семье повлияла на мой выбор	Менің отбасымның сол уақыттағы жағдайы таңдауыма әсер етті
26	Famous personalities who had the same specialization in that field affected my major selection	Знаменитые личности, которые имели ту же специализацию в этой области, повлияли на мой основной выбор	Осы салада маманданған танымал тұлғалар негізгі таңдауыма әсер етті
27	.My skills were major effect in my choice	Мои навыки оказали большое влияние на мой выбор	Менің қабілеттерім таңдауыма үлкен әсер етті

B. Appendix B

```
from sklearn.decomposition import PCA
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
plt.style.use('ggplot')
from sklearn.preprocessing import scale
import sklearn.datasets
from factor_analyzer import FactorAnalyzer

df= pd.read_excel("last.xlsx")
df.drop(['Choose your age group', 'Choose your Gender',
        'University Major/Specialization',
        'Your current university',
        'University GPA'],axis=1,inplace=True)

numCols=df.iloc[:,0:28]
numCorr=numCols.corr()
pd.DataFrame.from_records(numCorr)

from factor_analyzer.factor_analyzer import calculate_bartlett_
chi_square_value,p_value=calculate_bartlett_sphericity(df)
chi_square_value, p_value

from factor_analyzer.factor_analyzer import calculate_kmo
```

```
kmo_all,kmo_model=calculate_kmo(df)
kmo_model
import numpy as np
eVal_corr,eVec_corr=np.linalg.eig(numCorr)
idx_corr=eVal_corr.argsort()[::-1]
eVal_corr=eVal_corr[idx_corr]
eVec_corr=eVec_corr[:, idx_corr]
print(eVal_corr.round(3))
print(eVec_corr.round(3))

import numpy as np
U,S,V=np.linalg.svd(numCorr)
print(S)
print(V)

def flip_vector_sign(eVec):
    for i in range(eVec.shape[1]):
        if(eVec[:,i].sum()<0):
            eVec[:,i]=-1*eVec[:,i]
    return eVec

np.set_printoptions(linewidth=1000)
Vt=flip_vector_sign(V.T)
Vt.round(3)

exp=eVal_corr*100/np.sum(eVal_corr)
accSum=np.cumsum(exp)
pcNum=list(range(0,27))
data=np.array([pcNum,eVal_corr,exp,accSum])
eigenValues=pd.DataFrame(data.T,columns=['PC','Eigenvalue',
"% of Variance Exp","Cumulative %"])
eigenNumbers=eigenValues.copy()
format_mapping={'PC':'{:,.0f}','Eigenvalue':'{:,.3f}',
'% of Variance Exp':'{:.3f}%','Cumulative %':'{:.3f}%'}
```

```
for key,value in format_mapping.items():
    eigenValues[key]=eigenValues[key].apply(value.format)
eigenValues

import matplotlib.pyplot as plt
plt.figure(figsize=(6,6))
eachExp=eigenNumbers.iloc[:,2]
plt.bar(pcNum,accSum,width=0.5,color='cyan',
alpha=0.2,label="Cumulative %")
plt.plot(pcNum,eachExp,label="% of Varance Explained")
plt.plot(pcNum,eachExp,'ro',label='_nolegend_')
plt.xlabel("Principal Components")
plt.ylabel("% of Variance Explained")
plt.title("% of variance Explained by PCs",fontsize=16)
plt.legend(loc='upper left')
plt.show()

eVec_corr6=Vt[:,:6]
eVal_corr6=eVal_corr[:6]
loadings6=eVec_corr6*np.sqrt(eVal_corr6)
print(loadings6.round(3))

matcom=loadingsDf.T
matcom
from numpy import eye, asarray, dot, sum, diag
from numpy.linalg import svd
from numpy import eye, asarray, dot, sum, diag
from numpy.linalg import svd
def varimax(Phi, gamma = 1.0, q = 20, tol = 1e-6):
    p,k = Phi.shape
    R = eye(k)
    d=0
    for i in xrange(q):
    d_old = d
```

```
        Lambda = dot(Phi, R)
        u,s,vh = svd(dot(Phi.T, asarray(Lambda)**3 - (gamma/p) * do
            R = dot(u,vh)
            d = sum(s)
            if d_old!=0 and d/d_old < 1 + tol: break
        return dot(Phi, R)

l=varimax(matcom)
```

References

[1] A.S. Aldosary and S.A. Assaf. "Analysis of factors influencing the selection of college majors by newly admitted students". In: *Higher Education Policy* 9 (1996).

[2] M. Avellaneda and J. Lee. "Statistical arbitrage in the US equities market". In: *Quantitative Finance* 10.7 (2010), pp. 761–782.

[3] B. Beggs, J. Mullins, and T Taylor. "Distinguishing the factors influencing college students' choice of major". In: *College Student Journal* 42.2 (2008), pp. 381–395.

[4] J. Brown. "How are PCA and EFA used in language test and questionnaire development?" In: *Statistics* 14.2 (2010).

[5] A. Ciriaci D.and Muscio. "University choice, research quality and graduates' employability: Evidence from Italian national survey data". In: *European Educational Research Journal* 13.2 (2014), pp. 199–219.

[6] M. Daly. "Gender mainstreaming in theory and practice". In: *Social Politics: International Studies in Gender, State & Society* 12.3 (2005), pp. 433–450.

[7] D. DeMarie and P.A. Aloise-Young. "College students' interest in their major". In: *College Student Journal* 37.3 (2003), pp. 462–470.

[8] D. Fizer. "Factors affecting career choices of college students enrolled in agriculture". In: *A research paper presented for the Master of Science in Agriculture and Natural Science degree at The University of Tennessee, Martin* (2013).

[9] Galotti and M. Kathleen. "Making a" major" real-life decision: College students choosing an academic major." In: *Journal of Educational Psychology* 91.2 (1999), p. 379.

[10] S. Ghosh and J. Jintanapakanont. "Identifying and assessing the critical risk factors in an underground rail project in Thailand: a factor analysis approach". In: *International Journal of Project Management* 22.8 (2004), pp. 633–643.

[11] R. Hubbard and M. Bayarri. "Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing". In: *The American Statistician* 57.3 (2003), pp. 171–178.

[12] J. Jauregui. "Principal component analysis with linear algebra". In: *Philadelphia: Penn Arts & Sciences* (2012).

[13] H.F. Kaiser. "The application of electronic computers to factor analysis". In: *Educational and psychological measurement* 20.1 (1960), pp. 141–151.

[14] H. Li and P. Ralph. "Local PCA shows how the effect of population structure differs along the genome". In: *Genetics* 211.1 (2019), pp. 289–304.

[15] S. Lin and Y. Huang. "Development and application of a Chinese version of the short attitudes toward mathematics inventory". In: *International Journal of Science and Mathematics Education* 14.1 (2016), pp. 193–216.

[16] M. Lopez. "Estimation of Cronbach's alpha for sparse datasets". In: *Proceedings of the 20th Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ)*. 2007, pp. 151–155.

[17] C.A. Malgwi, M. A Howe, and P. A. Burnaby. "Influences on students' choice of college major". In: *Journal of Education for Business* 80.5 (2005), pp. 275–282.

[18] M. N. Marshall. "Sampling for qualitative research". In: *Family practice* 13.6 (1996), pp. 522–526.

[19] W.E. Martin and K. D Bridgmon. *Quantitative and statistical research methods: From hypothesis to results*. Vol. 42. John Wiley & Sons, 2012.

[20] P. McCullagh. "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

[21] A.N. Oppenheim. *Questionnaire design, interviewing and attitude measurement*. Bloomsbury Publishing, 2000.

[22] C. Papanastasiou and E. Papanastasiou. “Factors that influence students to become teachers”. In: *Educational Research and Evaluation* 3.4 (1997), pp. 305–316.

[23] K. Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[24] R.S. Sandman. “The Mathematics Attitude Inventory: Instrument and User’s Manual.” In: *Journal for research in Mathematics Education* 11.2 (1980), pp. 148–49.

[25] A. Sarwar and R Masood. “FACTORS AFFECTING SELECTION OF SPECIALIZATION BY BUSINESS GRADUATES.” In: *Science International* 27.1 (2015).

[26] S. Singh, M. Kapse, and J. Sonwalkar. “Factors which affect the career and subject preference of the female students of business schools”. In: *Journal of Women’s Entrepreneurship and Education* 1-2 (2011), pp. 89–107.

[27] D. Skowronek and L. Duerr. “The convenience of nonprobability: Survey strategies for small academic libraries”. In: *College & Research Libraries News* 70.7 (2009), pp. 412–415.

[28] L.I. Smith. *A tutorial on principal components analysis*. Tech. rep. 2002.

[29] B.G. Tabachnick, L.S. Fidell, and J. B. Ullman. *Using multivariate statistics*. Vol. 5. Pearson Boston, MA, 2007.

[30] C. Turan, S. Kadyrov, and D. Burissova. “An Improved Face Recognition Algorithm Based on Sparse Representation”. In: *2018 International Conference on Computing and Network Communications (CoCoNet)*. IEEE. 2018, pp. 32–35.

[31] A. K. Weller S.C .and Romney. *Systematic data collection*. Vol. 10. Sage publications, 1988.

[32] A. L. White et al. “Mathematical attitudes, beliefs and achievement in primary pre-service mathematics teacher education”. In: *Mathematics teacher education and development* 7.33-52 (2005).

[33] Sedat Yazici and Asli Yazici. “Students’ choice of college major and their perceived fairness of the procedure: evidence from Turkey”. In: *Educational Research and Evaluation* 16.4 (2010), pp. 371–382.