

**Federal State Autonomous Educational Institution of Higher Professional Education
National Research University
Higher School of Economics**

Final project for the course ‘Data Analysis’

Life Satisfaction of Russian citizens

Performed by
Glushkova Daria

Introduction

For a long time, the population's satisfaction with life was perceived as an important, but generally secondary indicator of the socio-economic development of society. Recently, however, not only abroad, but also in Russia, this indicator has begun to attract more and more attention.

Sociologists regularly conduct surveys, trying to find out how satisfied Russians are with their lives. Some survey results indicate a fairly high level of satisfaction with life, while others, on the contrary, that it is quite low. We also became interested in trying to assess the level of satisfaction with life in Russia by doing then on one of the most reliable bases that now exist in Russian science, and try to find out who is more satisfied with their lives, as well as what can be associated with this satisfaction.

We have set ourselves several hypotheses, which we will talk about in the course of the study and which we will either confirm or refute, depending on the results of the tests.

The main hypotheses are:

- 1) Women feel more satisfied with life than men;
- 2) Residents of cities are satisfied with life more than residents of villages and villages;
- 3) There is a significant impact of health status on the level of life satisfaction;
- 4) The "degree" of religiosity can significantly affect the level of satisfaction with life and believers are more satisfied with it.
- 5) People who feel respected feel satisfied with life
- 6) People who think they have a lot of power feel content with life.

Description of the research base and justification for the choice of the dependent variable

The empirical basis of the project was **the combined database of 1994-2018** of the results of surveys of **individuals within the framework** of the **Russian Monitoring of the Economic Situation and Health of the Population** (hereinafter referred to as "RLMS").

The data for the database was collected by the National Research University The Higher School of Economics and Demoscope LLC with the participation of the Population Center of the University of North Carolina at Chapel Hill and the Institute of Sociology of the Federal Research Sociological Center of the Russian Academy of Sciences as part of a longitudinal study, which is a series of annual nationwide representative surveys based on a probabilistic stratified multistage territorial sample, developed with the participation of the world's leading experts ¹.

The initial array consisted of 353,827 responses.

There are two questions in RLMS that can be used to assess the level of life satisfaction - "How satisfied are you with your life as a whole at the present time?" (**variable J65**) and "Tell me, please, are you happy?" (**variable M170**). Both questions have a five-level gradation. The most up-to-date RLMS data on the answer to the first question of 2018, the most up-to-date data in the public domain on measuring the level of happiness are presented in 2017. In 2017, the question about the level of satisfaction with respondents was also asked and the results of the survey are also open.

Nastina E.A. and Almakaeva A.M. in the article "The Role of the Level of Claims and Social Comparisons in the Determination of Life Satisfaction"² write that they also had a choice - to use the results of the answer to any one of these questions or to combine these two dimensions into an index. However, after estimating the correlation coefficient on the basis they used (an individual questionnaire of 2016), they found that "only 25% of the variation in life satisfaction is explained by level of happiness" and decided not to combine observations into an index, but to use the answers received to the question "How satisfied are you with your life as a whole at the present time?". We also decided to estimate the correlation coefficient on our data and made a similar conclusion.

¹ Joint database of 1994-2018 results of surveys of individuals of the RLMS of the Higher School of Economics. URL: <https://www.hse.ru/rlms/spss>. Access mode: free (accessed 16.05.2020)

² Nastina E.A. and Almakaeva A.M. The role of the level of claims and social comparisons in determining lifesatisfaction. URL: <https://cyberleninka.ru/article/n/rol-urovnya-prityazaniy-i-sotsialnyh-sravneniy-v-determinatsii-udovletvorennosti-zhiznyu> Mode dostup: free (accessed 16.05.2020)

In this regard, we decided to use one **variable J65** in our work, since it seems to us that at the level of perception, "happiness" may look more like for respondents a serious category of emotional content than "satisfaction", and therefore it is also more difficult for them to see achievable.

As a result, due to the fact that within the framework of the project we decided to investigate the level of life satisfaction of adults in Russia on the most relevant observations, the results obtained only in 2018 on the basis of the Main Observations were selected from the existing observations. Use of an adult questionnaire when interviewing respondents over 17 years old (drop if (**YEAR<2018**), drop if(**ADULT!=1**), **drop if (AGE<18)**).

Thus, 14,153 observations corresponding to the objectives of the study were obtained.

Selection of indicators of life satisfaction and variables to be used in the study

The problem of determining the factors influencing the level of life satisfaction has been occupying the scientific community for many years. There is no consensus, and different scientists use different approaches. Nevertheless, many, including, for example, Balatsky E.V.³, believe that the measurement should form the most complete set of factors that can potentially have an impact. We fully agree with this thesis, so our selection of variables was carried out carefully.

First of all, they identified the main "characteristics of life" - those aspects of a person's life that can describe him. These are:

- work;
- education;
- marital status;
- attitude to religion;
- financial situation;
- health;
- physical activity;
- bad habits;
- leisure and hobbies;
- emotional attitude to external circumstances.

³ Balatsky E.V. Life Satisfaction Factors: Measurement and Integral Indicators. URL: <https://cyberleninka.ru/article/n/factory-udovletvorennosti-zhiznyu-izmerenie-i-integralnye-pokazateli/viewer>. access mode: free (accessed 16.05.2020)

This was done in order to set a benchmark and not to miss important aspects in the selection. Next, in the questionnaire for adults, we searched for questions that would allow us to get an idea of these aspects of a person's life.

The result is a set of questions and variables presented in Table 1.

Aspect	Question (s)	Variable(s)	Measurement scale
Work	What is your main occupation currently, are you currently working, are you on paid or unpaid leave, or are you unemployed?	J1	Nominal (in the future will be converted in dichotomous)
	Tell me, please, how satisfied or dissatisfied you are with your work in general?	_v8	Nominal (in the future will be converted in ordinal)
	Tell me, please, to what extent are you satisfied or dissatisfied with the working conditions?	_v9	Nominal (in the future will be converted in ordinal)
	Tell me, please, how satisfied or dissatisfied you are with the remuneration of your work?	_v10	Nominal (later it will be converted to ordinal)
	Tell me, please, how satisfied or dissatisfied you are with the opportunities for your professional growth?	_v11	Nominal (later it will be converted to ordinal)
	What industry do you work in this job? * This refers to the main place of work	_v19	Nominal
	Do you have subordinates?	J6	dichotomous

Education	What is your highest level of education?	_v363	Nominal
	"Education (detailed)	EDUC	Nominal

	<i>The list of available answers to the second question is more in line with the needs of our study, however, in descriptive statistics we used EDUC</i>		
Marital status	Marital status in the current wave Do you have children, relatives or officially adopted?	MARST _v477	Nominal dichotomous
Attitude to religion	What is your attitude to religion?	_v451	Nominal (later it will be converted to ordinal)
Material situation	Do you or your family have the opportunity to eat meat, chicken or fish every other day if you wish? Do you or your family have the opportunity, if you wish, to improve your living conditions - to buy a room, apartment, house? Do you or your family have the opportunity, if you wish, to save money for large purchases - a car, a cottage? Do you or your family have the opportunity to spend a vacation abroad with the whole family if you wish ? Do you or your family have the opportunity, if you wish, to spend a vacation with the whole family at a Russian resort? How much money have you received in the last 30 days at your main place of work after taxes? Tell me, please, how satisfied are you	J721637 J721631 J721633 J721634 J721636 J10	Dichotomous Dichotomous Dichotomous Dichotomous Dichotomous Interval

	with your financial situation at the present time?	_v156	Nominal
--	--	-------	---------

			(later will be converted to ordinal)
Health	Have you had any health problems in the last 30 days? Tell me, please, how do you assess your health?	L5 M3	dichotomous Nominal (later will be converted to ordinal)
Physical activity	Which of the descriptions best suits your physical education classes? Please do not consider Workload	M114	Nominal (later will be converted to ordinal)
Bad habits	Do you currently smoke ? Have you consumed alcoholic beverages in the last 30 days?	M71 M80	dichotomous dichotomous
Leisure and hobbies	<i>In the 2018 Adult Questionnaire, questions related to this aspect of life are asked by us were not found</i>	-	-
Emotional attitude to external circumstances	Do you think that most people can be trusted or that you should always be careful when dealing with people? Now imagine , please, a ladder of 9 steps, where on the lower step there are completely disenfranchised, and on the highest - those who have great power. Which of the nine steps are you on today? And another staircase of 9 steps, where on the lower step there are people who are not respected at all, and on the highest - those who are very respected.	J206 J63 J64	Nominal (later will be converted to ordinal) Nominal (later will be converted to ordinal) Nominal (later it will be converted to ordinal)

	Which of the nine Are you personally on the steps today?		
--	---	--	--

Table 1. Variables selected for the study

We believe that the set of variables we have identified will be sufficient to conduct the study.

We also decided to use the following categories in the analysis (Table 2)

Variable description	Variable	Type of measurement scale
Occupation	J90	Nominal
Floor	H5	Dichotomous
Age	AGE	Interval
Region of residence	REGION	Nominal
Type of locality	STATUS	Nominal

Table 2. Categories selected for the study We assume that the distribution of respondents into these categories may also arouse interest.

Variable transcoding

Quite a few variables have been selected that are nominal, but can be turned into ordinal variables after conversion. It was decided to carry out the transformation, since it is easier to work with ordinal variables than with nominal variables.

Variable(s)	Former value	New Value
_v8, _v9, _v10, _v11,_v156	Fully satisfied	1
	Rather satisfied	2
	Yes and no	3
	Rather not satisfied (Not very satisfied)	4
	Not at all satisfied	5
_v451	You are a believer	1
	You are, rather, a believer than an unbeliever person	2
	You are an unbelieving person	3
	You are an unbeliever	4
	You are an atheist	5
M3	Very good	1
	Good	2
	Average, not good, but not bad either	3
	Bad	4
	Really bad	5

M114	Light exercise for rest less than three times a week	2
	Exercise of moderate or high severity less than three times a week	3
	High-weight exercise at least three times a week for 15 minutes and more	4
	Daily physical activity at least At least, 30 minutes a day	6
	Daily physical activity of less than 30 minutes a day	5
	He does not engage in physical education	1
J206	Most people can be trusted	1
	In relationships with people, you always have to be Careful	2
	Both, depending on the person, the conditions	3
J63, J64	Lowest level	1
	The Highest Level	9

Table 3. Transcoding variables from nominal to ordinal variables

The values of some nominal variables, as well as all dichotomous ones, were recoded from text to numeric (Table 4). Transcoding was carried out in an Excel file, so there are no changes to the names of the variables.

Variable(s)	Former value	New meaning
MARST	Never married	1
	You are married	2
	Living together, but not registered	3
	Divorced and unmarried	4
	Widower (widow)	5
	Officially registered, but not together Live	6
H5	Male	1

	Female	2
J1	You are currently working	1
	You are on any other paid leave	
	You are on unpaid leave	
	You are on maternity leave or to care for a child under 3 years old	2
	Or you don't have a job right now	
J6, _v477,	Yes	1
	No	0
J721637,		
J721631,		
J721633,		
J721634,		
J721636, L5,		
M71, M80		

Table 4. The result of transcoding text values into numeric values

Descriptive data analysis

To "get acquainted" with the resulting sample, we will conduct a descriptive analysis of the data.

It should be noted that in the RLMS there are three types of answer options for each question, which do not carry a meaningful load: "I find it difficult to answer", "No answer", "Refusal to answer". We have cleaned our data of these answers because they affect the percentage of variants, but do not conclude in nothing on a semantic level.

Let's construct frequent distributions for some of the values we have chosen. This will help us get a general idea of our sample.

tabulate H5

Пол респондента	Freq.	Percent	Cum.
МУЖСКОЙ	5,909	41.75	41.75
ЖЕНСКИЙ	8,244	58.25	100.00
Total	14,153	100.00	

Table 5. Distribution of respondents relative to their gender

Table 5 shows the predominance of the female audience of respondents: about 58% of the questionnaires contained in the database are women's questionnaires, while the "male" questionnaires are about 42%. Recall that the RLMS positions its sample as representative of the population of the Russian Federation. According to Rosstat at the beginning of 2019

There were 68.1 million men and 78.6 million women living in the country.⁴ The percentage of men to women is 46%. If we consider that respondents under the age of 18 were excluded from our sample, and boys are traditionally born in the country more than girls, but the percentage of child mortality among boys. On average, higher than among girls, we can say that the ratio of men to women among respondents generally roughly coincides with the observed ratio of men and women in the country and the differences can be considered Irrelevant.

Now let's find out the age of the respondents. To do this, first of all, we will distribute the age values stored in the **AGE** variable into groups.

To do this, create a new variable of string type **AGE_cat** and distribute the values of all ages in it into groups 18-25, 26-35, 36-45, 46-55, 56-65, 66-75, from 76. generate str AGE_cat = "hello world"

replace AGE_cat = "18-25" if AGE < 26

replace AGE_cat = "26-35" if (AGE > 25 & AGE < 36)

replace AGE_cat = "36-45" if (AGE > 35 & AGE < 46)

replace AGE_cat = "46-55" if (AGE > 45 & AGE < 56)

replace AGE_cat = "56-65" if (AGE > 55 & AGE < 66)

replace AGE_cat = "66-75" if (AGE > 65 & AGE < 76)

replace AGE_cat = "from 76" if AGE > 75

```
tabulate AGE_cat
```

AGE_cat	Freq.	Percent	Cum.
18-25	1,451	10.25	10.25
26-35	2,835	20.03	30.28
36-45	2,731	19.30	49.58
46-55	2,189	15.47	65.05
56-65	2,363	16.70	81.74
66-75	1,426	10.08	91.82
from 76	1,158	8.18	100.00
Total	14,153	100.00	

Table 6. Distribution of respondents group by age

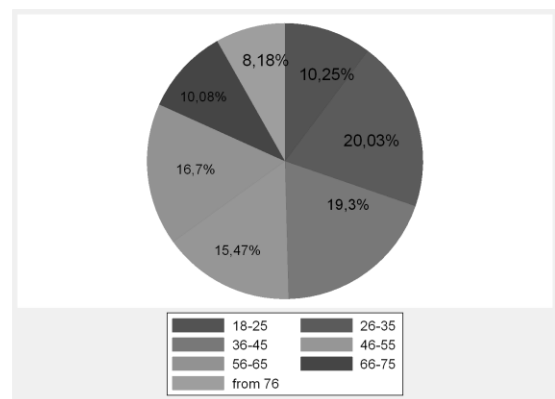


Figure 1. Distribution of respondents by age group (graph **pie, over(AGE_cat)**)

From Table 6 and Figure 1, we can see that approximately half (49.58%) of respondents are under 46 years of age. The most numerous are the groups 26-35 (20.03%) and 36-45 (19.30%). In general, the groups turned out to be relatively equal in fullness,

⁴ Rosstat named the ratio of the number of men and women in Russia. URL: <https://ria.ru/20200422/1570386332.html>, access mode: free (accessed 17.05.2020)

Therefore, in the future, if there is a need for analysis by age groups, this distribution can be used without making additional changes to it.

`. tabulate STATUS`

STATUS	Freq.	Percent	Cum.
Gorod	3,727	26.33	26.33
Oblastnoj tsentr	5,984	42.28	68.61
PGT	982	6.94	75.55
Selo	3,460	24.45	100.00
Total	14,153	100.00	

Table 7. Distribution of respondents by type of locality

Table 7 shows the distribution of respondents by type of locality in which they lived at the time of answering the questions. Most live in cities, which are regional centers. Nevertheless, about a quarter of respondents are from rural areas and about 7% are from urban-type settlements. Such a distribution, in principle, allows us to talk about representativeness, since the level of urbanization of the country is approximately at around 74.5%⁵.

`. tabulate MARST`

MARST	Freq.	Percent	Cum.
1	2,086	14.77	14.77
2	7,205	51.01	65.78
3	1,728	12.23	78.01
4	1,262	8.93	86.95
5	1,786	12.64	99.59
6	58	0.41	100.00
Total	14,125	100.00	

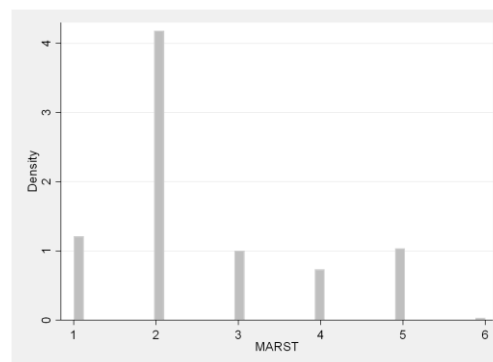


Table 8 and Figure 2. Distribution of respondents by marital status

(never married - 1, are in a registered marriage - 2, live together, but not registered - 3, divorced and unmarried - 4, widower (widow) - 5, officially registered, but do not live together – 6) (**histogram MARST**)

Based on Table 8 and Figure 2, it is possible to draw conclusions about the marital status of the respondents. Half of them are married (51.01%).

⁵ Urbanization in Russia. Rosinfostat. URL: <https://rosinfostat.ru/urbanizatsiya-v-rossii/>, access mode: free (accessed 18.05.2020)

. tabulate J90

J90	Freq.	Percent	Cum.
Domashnyaya khozyajka, ukhazhivaete z..	522	3.70	3.70
Fermer	6	0.04	3.75
Nakhodites' v dekretnom otpuske	45	0.32	4.07
Nakhodites' v ofitsial'nom otpuske po..	163	1.16	5.22
Ne rabotaete po sostoyaniyu zdorov'ya..	161	1.14	6.37
Pensioner i ne rabotaete	3,971	28.18	34.55
Predprinimatel'	207	1.47	36.02
Rabotaete na predpriyatii, v organiza..	6,932	49.20	85.22
Rabotaete ne na predpriyatii, v organ..	602	4.27	89.49
Student dnevnogo VUZa	347	2.46	91.95
Student dnevnogo tekhnikuma, kolledzha	128	0.91	92.86
Uchenik srednej shkoly, PTU	53	0.38	93.24
Vremenno ne rabotaete po drugim prich..	341	2.42	95.66
Vremenno ne rabotaete po drugim prich..	612	4.34	100.00
Total	14,090	100.00	

Table 9. Distribution of respondents by occupation

We were also interested in finding out the main occupation of the respondents. Table 9 indicates that the vast majority of them work in an enterprise/organization (49.2%). There are also quite a lot of pensioners among the respondents - more than a quarter.

. tabulate DIPLOM

DIPLOM	Freq.	Percent	Cum.
nezakonchennoe srednee obrazovanie (7..	613	4.34	4.34
nezakonchennoe srednee obrazovanie (7..	1,096	7.75	12.09
okonchil 0 - 6 klassov	232	1.64	13.73
zakonchennoe srednee obrazovanie	4,265	30.16	43.89
zakonchennoe srednee spetsial'noe obr..	3,860	27.30	71.19
zakonchennoe vysshee obrazovanie i vy..	4,073	28.81	100.00
Total	14,139	100.00	

Table 10. Distribution of respondents by level of education

A cross-section of the most common groups denoting the level of education showed that the groups of those who have completed higher education, those who have secondary special education and those who have completed secondary education are approximately the same education.

We were also interested in finding out the average value of wages among our respondents and the median level of wages. Our variable may contain outliers, so we initially used the tab J10 command to derive the frequency distribution and estimate their presence.

tab J10

J10	Freq.	Percent	Cum.
0	218	2.87	2.87
50	8	0.11	2.98
57	1	0.01	2.99
60	1	0.01	3.01
850	1	0.01	3.02
1200	6	0.08	3.10
1380	2	0.03	3.12
1400	1	0.01	3.14
1500	2	0.03	3.16
1800	1	0.01	3.18
2000	5	0.07	3.24
2400	1	0.01	3.26
2500	2	0.03	3.28
2600	1	0.01	3.30
2800	1	0.01	3.31
2900	1	0.01	3.32
3000	9	0.12	3.44
3070	1	0.01	3.45
3200	2	0.03	3.48
3260	1	0.01	3.49
3500	3	0.04	3.53
3646	1	0.01	3.55
3700	2	0.03	3.57

Table 11. Part of the frequency distribution

Since the minimum wage in the country is at the level of 12,000, we decided to exclude all values less than 6000 (minimum wage at half rate). We decided that we would not need such observations further in the entire study, so we got rid of them radically with the drop if J10<6000 command. 352 observations were excluded from the sample .

Then, finally, we moved on to the output of the average value. We did this with the help of the sum J10, d command

. sum J10, d				
J10				
	Percentiles	Smallest		
1%	6700	6000		
5%	9800	6000		
10%	11000	6000	Obs	7,233
25%	15000	6000	Sum of Wgt.	7,233
50%	25000		Mean	28337.88
		Largest	Std. Dev.	20504.26
75%	35000	267000		
90%	50000	300000	Variance	4.20e+08
95%	60000	350000	Skewness	4.427147
99%	100000	400000	Kurtosis	46.75139

Table 12. The average salary of all respondents received by them in the last 30 days (after taxes) at the main place of work

Table 12 shows that among working respondents, 7233 agreed, the average salary was 28,337 rubles. The median salary is 25,000 rubles.

For the sake of satisfying the research interest, we also calculated the average salary by age groups (tabstat J10, statistics(mean) by(AGE_cat) (Table 13)

Summary for variables: J10
by categories of: AGE_cat

AGE_cat	mean
18-25	24990.57
26-35	29978
36-45	30688.25
46-55	27446.86
56-65	23954.26
66-75	20498.1
from 76	13352.94
Total	28337.88

Table 13. Average salary of all respondents received by them in the last 30 days (after taxes) at the main place of work, by age groups

Table 13 shows that there are no significant differences between the wages of people from 18 to 65 years of age (the period of active employment).

Exploring Relationships Between Variables: Contingency Tables

Estimating the level of wages by age groups, we smoothly approached the assessment of relationships and relationships between various variables.

The nature of the relationship between two variables can be visualized using contingency tables. However, we note that the analysis of contingency tables reveals the presence/absence of statistical, rather than causal, dependencies. Also note that the construction of contingency tables is not very advisable without assessing the significance, so we will do it for each table. To do this, we will use the Xi-square, since it basically has no restrictions on the type of variables. If the P-value tends to zero, i.e. <0.05 , then we can conclude that there is a statistically significant relationship between the variables, since the significance is less than 0.05.

In the course of our study of life satisfaction of the population, it is necessary to identify the factors influencing it. It is natural to assume that the higher the level of income of a person, the higher the degree of his satisfaction with life. This hypothesis is very easy to test by constructing a two-dimensional contingency table.

To begin with, let's distribute the salary values into several groups to make the assessment easier.

To do this, created a new variable of string type SALARY_cat and distributed in it the values of wages per group up to 15,000, from 15,001 to 25,000 , from 25,001 to 40,000, from 40,001 to 60,000, from 60,001 to 80,000, from 80,001 rubles. generate str SALARY_cat = "hello world"

```
replace SALARY_cat = "<15 001" if (J10 >5999 & J10 <15001)
replace SALARY_cat = "15 001-25 000" if (J10 >15000) & (J10 < 25001)
replace SALARY_cat = "25 001-40 000" if (J10 >25000) & (J10 < 40001)
replace SALARY_cat = "40 001-60 000" if (J10 >40000) & (J10 < 60001)
replace SALARY_cat = "60 001-80 000" if (J10 >60000) & (J10 < 80001)
replace SALARY_cat = "from 80 001" if (J10 >80000)
replace SALARY_cat = "no group" if J10 ==-1
```

It should be noted that initially, when we proceeded to the creation of the "from 80,001" group, Stata also included respondents whose salary level was not available. We have found a solution to this problem by filling the values of blank cells in Excel with a value of -1. The resulting frequency distribution of respondents by the level of their wages is presented in Table 14.

```
. tabulate SALARY_cat
```

SALARY_cat	Freq.	Percent	Cum.
15 001-25 000	2,289	16.59	16.59
25 001-40 000	1,980	14.35	30.93
40 001-60 000	829	6.01	36.94
60 001-80 000	183	1.33	38.27
<15 001	1,826	13.23	51.50
from 80 001	126	0.91	52.41
no group	6,568	47.59	100.00
Total	13,801	100.00	

Table 14. Frequency distribution of respondents according to the level of their wages

Now that the SALARY_cat variable is ready, we can move on to building the contingency table.

SALARY_cat	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
15 001-25 000	639 587.7	342 409.4	153 172.7	1,051 961.4	91 144.8	2,276 2,276.0
25 001-40 000	488 510.8	256 355.8	180 150.0	1,009 835.5	45 125.9	1,978 1,978.0
40 001-60 000	154 212.8	57 148.2	135 62.5	464 348.1	14 52.4	824 824.0
60 001-80 000	34 47.3	17 32.9	30 13.9	99 77.3	3 11.6	183 183.0
<15 001	558 469.7	352 327.2	104 138.0	688 768.3	117 115.7	1,819 1,819.0
from 80 001	20 32.5	11 22.7	31 9.6	63 53.2	1 8.0	126 126.0
no group	1,654 1,686.2	1,436 1,174.7	409 495.4	2,428 2,758.2	603 415.5	6,530 6,530.0
Total	3,547 3,547.0	2,471 2,471.0	1,042 1,042.0	5,802 5,802.0	874 874.0	13,736 13,736.0

Pearson chi2(24) = 735.1263 Pr = 0.000
likelihood-ratio chi2(24) = 738.0017 Pr = 0.000

Table 15. Frequent distribution of respondents by salary level, reflecting life satisfaction (tabulate SALARY_cat J65, chi2 expected lrchi2)

From Table 15, we can see that our hypothesis is not reflected in reality – respondents with a salary below the median make up the majority of those who answered that they are "rather satisfied" with their lives than not. Also, people with a salary of 25001-40000 are "rather satisfied". Contrary to expectations, people with a salary above the median can be either "completely satisfied with life", as well as "not at all satisfied."

```
. tab J65 STATUS,chi2
```

J65	STATUS				Total
	Gorod	Oblastn..	PGT	Selo	
I da, i net	964	1,409	241	933	3,547
Ne ochen' udovletvo..	663	1,025	139	644	2,471
Polnost'yu udovletv..	284	477	93	188	1,042
Skoree udovletvoreny	1,454	2,560	440	1,348	5,802
Sovsem ne udovletvo..	236	349	47	242	874
Total	3,601	5,820	960	3,355	13,736

Pearson chi2(12) = 70.7544 Pr = 0.000

Table 15. The conjugation of the level of satisfaction with life and type locality

We will also construct a table of contingency between the level of life satisfaction and the type of settlement in which the respondents live. It tells us that most of those living in the regional center are "more satisfied" with life than not. In general, we also cannot claim that there is a clear relationship between the type of settlement and the level of satisfaction.

MARST	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
1	529 526.7	343 366.8	176 154.4	863 861.4	128 129.7	2,039 2,039.0
2	1,751 1,797.6	1,067 1,251.9	606 527.0	3,196 2,939.9	339 442.7	6,959 6,959.0
3	455 430.6	278 299.9	115 126.2	718 704.2	101 106.0	1,667 1,667.0
4	336 316.2	312 220.2	45 92.7	388 517.1	143 77.9	1,224 1,224.0
5	456 454.9	455 316.8	90 133.3	605 743.9	155 112.0	1,761 1,761.0
6	14 15.0	11 10.4	6 4.4	21 24.5	6 3.7	58 58.0
Total	3,541 3,541.0	2,466 2,466.0	1,038 1,038.0	5,791 5,791.0	872 872.0	13,708 13,708.0

Pearson chi2(20) = 366.2913 Pr = 0.000
likelihood-ratio chi2(20) = 356.5416 Pr = 0.000

Table 16. The conjugation of the level of life satisfaction and marital status

From Table 16, we can see that the vast majority of those living together, but not registered, are "rather satisfied" with life. Among those who are completely satisfied, the number of those who live together, but is not registered, also prevails.

. tab AGE_cat J65, chi2

AGE_cat	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
18-25	339	166	193	665	54	1,417
26-35	679	383	252	1,302	111	2,727
36-45	725	423	217	1,110	146	2,621
46-55	563	399	113	878	157	2,110
56-65	599	480	114	914	191	2,298
66-75	370	368	69	495	112	1,414
from 76	272	252	84	438	103	1,149
Total	3,547	2,471	1,042	5,802	874	13,736

Pearson chi2(24) = 396.4792 Pr = 0.000

Table 17. Conjugation of the level of life satisfaction and age

Table 17 tells us that in all age categories, the variant is in the lead "rather satisfied" than not. In second place in all groups "yes and no",

```
. tab J1 J65, chi2
```

J1	J65					Total
	I da, i...	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
1	1,921	1,057	626	3,387	284	7,275
2	1,624	1,414	415	2,413	589	6,455
Total	3,545	2,471	1,041	5,800	873	13,730

Pearson chi2(4) = 341.5961 Pr = 0.000

Table 18. The relationship between the level of life satisfaction and the presence/absence of work

Table 18 did not surprise us with its results. The most popular option in all categories is still "rathersatisfied", in second place is still "yes and no".

```
. tab _v8 J65, chi2
```

_v8	J65					Total
	I da, i...	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
1	167	85	251	545	18	1,066
2	1,004	466	338	2,198	93	4,099
3	597	329	51	530	102	1,609
4	143	151	11	160	43	508
5	33	30	3	37	26	129
Total	1,944	1,061	654	3,470	282	7,411

Pearson chi2(16) = 995.4181 Pr = 0.000

Table 19. The conjugation of the level of life satisfaction and the level of job satisfaction

The variable _v8 stores respondents' answers to the question "Tell me, please, how satisfied or dissatisfied you are with your work in general"? Among those who are "rather satisfied with their jobs," an unexpectedly large proportion did not answer that they were also "rather satisfied with life," but chose the "yes and no" option. Among those who I am not at all satisfied with life more than those who, when asked about work, answered "yes and no." Among those who are "not very satisfied with life" there are more of those who answered that they are "rather satisfied with work." These results suggest that work is not the main factor in life satisfaction for many people.

. tab J6 J65, chi

J6	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
0	1,640	913	501	2,748	255	6,057
1	321	165	158	759	34	1,437
Total	1,961	1,078	659	3,507	289	7,494

Pearson chi2(4) = 54.1880 Pr = 0.000

Table 20. The conjugation of the level of life satisfaction and the presence of subordinates

Table 20 suggests that the vast majority of those who do not have subordinates, as well as those who do not, are "rather satisfied" with their work.

. tab _v451 J65, chi2

_v451	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
1	982	664	410	1,741	257	4,054
2	1,667	1,211	415	2,799	369	6,461
3	506	328	108	673	114	1,729
4	197	132	55	289	80	753
5	83	77	26	152	32	370
Total	3,435	2,412	1,014	5,654	852	13,367

Pearson chi2(16) = 112.5323 Pr = 0.000

Table 21. Conjugation of the level of life satisfaction and attitude to religion

Table 21 shows interesting results. We hypothesized that there would be more believers who are "completely satisfied with life" than others. However, this option is only the fourth out of five in popularity. However, among atheists, the proportion of those who are completely satisfied with life is lower than among believers.

. tab M3 J65

M3	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
1	42	14	36	118	7	217
2	919	500	492	2,324	119	4,354
3	2,094	1,342	443	2,849	412	7,140
4	456	536	54	460	246	1,752
5	20	61	13	38	86	218
Total	3,531	2,453	1,038	5,789	870	13,681

Table 22. The conjugation of the level of life satisfaction and self-assessment of health

Table 22 tells us that among those who rated their health as "poor", the answer option "not very satisfied with life" prevails. Perhaps health plays a role

a significant role in the formation of attitudes towards life satisfaction. However , this hypothesis also needs to be tested.

. tab J65 J206

J65	J206			Total
	1	2	3	
I da, i net	386	1,552	1,554	3,492
Ne ochen' udovletvo..	294	1,177	960	2,431
Polnost'yu udovletv..	255	440	341	1,036
Skoree udovletvoreny	1,071	2,478	2,207	5,756
Sovsem ne udovletvo..	56	445	350	851
Total	2,062	6,092	5,412	13,566

Table 23. The conjugation of the level of life satisfaction and the willingness to trust people

Table 23 shows that among those who believe that "one should always be careful in dealing with people" and "both , depending on the person, conditions " in almost Equally, the answer "yes and no" prevails.

. tab J64 J65

J64	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
0	187	149	26	153	79	594
1	6	6	0	6	23	41
2	34	37	6	26	44	147
3	133	146	40	127	94	540
4	298	246	46	299	107	996
5	774	520	174	881	191	2,540
6	674	440	127	1,098	135	2,474
7	802	516	200	1,614	105	3,237
8	479	316	258	1,130	74	2,257
9	160	95	165	468	22	910
Total	3,547	2,471	1,042	5,802	874	13,736

Table 24. The conjugation of the level of life satisfaction and self-esteem of the level of respect in society

Table 24 shows that among those who believe that they are in society at the 9th level, there are more "rather satisfied with life".

. tab J63 J65

J63	J65					Total
	I da, i..	Ne oche..	Polnost..	Skoree ..	Sovsem ..	
0	111	64	24	107	14	320
1	158	209	65	203	197	832
2	426	467	96	464	249	1,702
3	820	672	172	999	231	2,894
4	798	540	156	1,098	99	2,691
5	818	350	286	1,507	63	3,024
6	285	106	113	811	13	1,328
7	113	55	74	414	3	659
8	14	6	42	105	2	169
9	4	2	14	94	3	117
Total	3,547	2,471	1,042	5,802	874	13,736

Table 25. The conjugation of the level of life satisfaction and self-esteem of the level of power

J63 contains the answer to the question "Now imagine, please, a ladder of 9 steps, where on the lower step there are completely disenfranchised, and on the highest - those who have great power. Which of the nine steps are you on today??" . Among those who have placed themselves on the highest level, the number of those who are "rather satisfied " with life prevails. Among those who are at the lowest (first) - the number of those who are "not very satisfied", but only a few answers are less for the option "rather satisfied" and at the same time, in third place, the answer is "not at all satisfied." In general, we can say that this group of people who are dissatisfied with life more.

The study of relationships between variables: correlation coefficients

The study of the contingency table, which describes the joint two-dimensional distribution of variables, is the first step in studying the relationship between two variables. However, it is not possible to evaluate or verify cause-and-effect relationships. To do this, you need correlation coefficients.

Since our variable J65, which denotes the degree of life satisfaction, is ordinal (rank), we need Spearman correlation coefficients (if the second variable is interval or ordinal (rank) or Kendall (if the second variable is also ordinal (rank)). Note that the use of the Kendall coefficient is welcome in samples in which there are outliers. We will use the Pepperman correlation coefficient, because there are no outliers in ordinal scales on pre-cleaned data.

Let's evaluate the pairing of J65 and _v8 (Tell me, please, how satisfied or dissatisfied you are with your work as a whole?). Both scales are ordinal (rank).

Since J65 has remained a nominal scale until now, it's time to recode it to ordinal J65_new by creating a new variable, which we did.

```
. spearman J65_new _v8

Number of obs =      7751
Spearman's rho =      0.3192

Test of Ho: J65_new and _v8 are independent
Prob > |t| =      0.0000
```

Table 26. The result of calculating the correlation coefficient for the variables denoting the level of life satisfaction and the level of job satisfaction

Table 26 shows the result of calculating the correlation coefficient for the variables denoting the level of life satisfaction and the level of job satisfaction. P-value indicates that the existing relationship is significant (because the P-value < 0.05). The Spearman coefficient is approximately 0.3 and indicates that there is a weak direct correlation between the variables. Therefore, we cannot say that these variables strongly influence each other. We will evaluate other factors.

```
. spearman J65_new _v451

Number of obs =    13686
Spearman's rho =     0.0588

Test of Ho: J65_new and _v451 are independent
Prob > |t| =     0.0000
```

Table 27. The result of calculating the correlation coefficient for variables denoting the level of life satisfaction and the level of religiosity Let us evaluate the relationship between the level of life satisfaction and attitude towards religion.

Based on Table 27, we see that there is practically no connection between the subjects.

```
. spearman J65_new _v156

Number of obs =    13701
Spearman's rho =     0.4432

Test of Ho: J65_new and _v156 are independent
Prob > |t| =     0.0000
```

Table 28. The result of calculating the correlation coefficient for the variables denoting the level of life satisfaction and the degree of satisfaction

Health

Table 28 tells us that there is a weak direct meaningful relationship.

```
. spearman J65_new AGE

Number of obs =    14082
Spearman's rho =     0.1510

Test of Ho: J65_new and AGE are independent
Prob > |t| =     0.0000
```

Table 29. The result of calculating the correlation coefficient for variables denoting the level of life satisfaction and age

Let's evaluate the relationship between the level of life satisfaction and age. Based on Table 29, we see that there is a weak direct correlation between the variables.

Analysis of variance

As part of the study, we are also interested in testing several hypotheses about the inequality of the values of the variable taken as a basis in different groups. To do this, tests are carried out to compare the averages.

Since in our case the dependent variable J65, which reflects the level of life satisfaction, is ordinal (rank), we should use nonparametric tests.

First of all, we will evaluate the average values of the level of life satisfaction of men and women. To do this, we will conduct a Mann-Whitney test.

ranksum J65_new, by (H5)

```
. ranksum J65_new, by ( H5 )

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

H5	obs	rank sum	expected
1	5875	40260086	41368813
2	8207	58898317	57789591
combined	14082	99158403	99158403


```
unadjusted variance    5.659e+10
adjustment for ties    -5.654e+09
-----
adjusted variance      5.093e+10

Ho: J65_new(H5==1) = J65_new(H5==2)
      z =  -4.913
      Prob > |z| =  0.0000
```

Table. Assessment of the average values of the level of life satisfaction of men (1) and women (2)

First of all, let's say that the result obtained is significant ($\text{Prob} > |z| = 0.00$, which is less than 0.05). Now let's evaluate the result itself. We can see that the rank sum is significantly higher in women (line "2") than in men (line "1"). Recall that the J-65_new ranks go from "more positive" to "less positive". And the highest - 5 is an indicator of lower life satisfaction. Based on this, you can make a statement about the fact that women on average are less satisfied with life than men.

```
. ranksum J65_new, by ( J1 )
```

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

J1	obs	rank sum	expected
1	7595	49619587	53457408
2	6481	49454339	45616519
combined	14076	99073926	99073926

```

unadjusted variance    5.774e+10
adjustment for ties    -5.770e+09
-----
adjusted variance      5.197e+10

Ho: J65_new(J1==1) = J65_new(J1==2)
    z = -16.834
Prob > |z| = 0.0000

```

Table. Assessment of the average values of the level of life satisfaction of workers (1) and non-workers (0)

Now let's check the averages in the group of those who work and those who don't. We see that those who work are less satisfied with life than those who are not. Perhaps this is due to the fact that there are quite a lot of pensioners in the sample.

```
. ranksum J65_new, by ( J6 )
```

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

J6	obs	rank sum	expected
0	6317	25362220	24753165
1	1519	5343146	5952201.5
combined	7836	30705366	30705366

```

unadjusted variance    6.267e+09
adjustment for ties    -7.817e+08
-----
adjusted variance      5.485e+09

Ho: J65_new(J6==0) = J65_new(J6==1)
    z = 8.224
Prob > |z| = 0.0000

```

Table. Assessment of the average values of the level of life satisfaction of those who have subordinates (1) and those who do not have them (0)

Those who do not have subordinates are less satisfied with life than those who do.

```
. ranksum J65_new, by ( _v477)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

_v477	obs	rank sum	expected
0	3074	20578351	21640960
1	11005	78537810	77475200
combined	14079	99116160	99116160

```

unadjusted variance  3.969e+10
adjustment for ties  -3.966e+09
-----
adjusted variance    3.573e+10

Ho: J65_new(_v477==0) = J65_new(_v477==1)
      z =  -5.622
Prob > |z| =  0.0000

```

Table. Assessment of the average values of the level of life satisfaction of those who have children (1) and those who do not (0)

Those who have children are less satisfied with life than those who do not.

```
. ranksum J65_new, by ( J721631)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

J721631	obs	rank sum	expected
0	12448	88320528	85853856
1	1345	6809793.5	9276465
combined	13793	95130321	95130321

```

unadjusted variance  1.925e+10
adjustment for ties  -1.913e+09
-----
adjusted variance    1.733e+10

Ho: J65_new(J721631==0) = J65_new(J721631==1)
      z =  18.736
Prob > |z| =  0.0000

```

Table. Assessment of average values of the level of life satisfaction those who have the opportunity to afford to improve their living conditions (1) and those who do not (0)

Those who have the opportunity to improve their living conditions are more satisfied with life than those who do not.

```
. ranksum J65_new, by ( L5)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

L5	obs	rank sum	expected
0	9035	59720462	63321798
1	4981	38510674	34909339
combined	14016	98231136	98231136

```

unadjusted variance    5.257e+10
adjustment for ties    -5.263e+09
-----
adjusted variance      4.730e+10

Ho: J65_new(L5==0) = J65_new(L5==1)
      z = -16.558
Prob > |z| = 0.0000

```

Table. Assessment of the average values of the level of life satisfaction of those who have had health problems in the last 30 days (1), and those who did not have them (2)

The table shows us that those who did not have problems, on average, are more satisfied with life.

Factor analysis

Factor analysis is a procedure by which a large number of variables related to available observations are reduced to a smaller number of independent influencing quantities, called factors:

- Variables that are strongly correlated with each other are combined into one factor.
- Variables from different factors are weakly correlated with each other.

We try to do a factor analysis. To do this, let's take dichotomous variables from our database that answer the questions:

- Do you have subordinates?
- Do you or your family have the opportunity to eat meat, chicken or fish every other day if you wish?
- Do you or your family have the opportunity, if you wish, to improve your living conditions - to buy a room, apartment, house?
- Do you or your family have the opportunity, if you wish, to save money for large purchases - a car, a cottage?
- Do you or your family have the opportunity to spend a vacation abroad with the whole family if you wish?
- Do you or your family have the opportunity, if you wish, to spend a vacation with the whole family at a Russian resort?

- Have you had any health problems in the last 30 days?
- Do you currently smoke ?
- Have you consumed alcoholic beverages in the last 30 days?

The computer selected 1973 observations from the respondents' questionnaires, selecting only those that had answers to all these questions.

```
. factor J721637 J721631 J721632 J721633 J721634 J721636 J721635 L5 M80 M71 J6 J65_new
> cf
(obs=1,973)
```

```
Factor analysis/correlation          Number of obs   =    1,973
Method: principal-component factors   Retained factors =     4
Rotation: (unrotated)                Number of params =    42
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	3.41216	2.26055	0.2843	0.2843
Factor2	1.15161	0.10060	0.0960	0.3803
Factor3	1.05101	0.03457	0.0876	0.4679
Factor4	1.01644	0.10419	0.0847	0.5526
Factor5	0.91225	0.05851	0.0760	0.6286
Factor6	0.85374	0.02121	0.0711	0.6998
Factor7	0.83253	0.07131	0.0694	0.7691
Factor8	0.76122	0.04543	0.0634	0.8326
Factor9	0.71579	0.20069	0.0596	0.8922
Factor10	0.51509	0.05078	0.0429	0.9352
Factor11	0.46432	0.15048	0.0387	0.9738
Factor12	0.31383	.	0.0262	1.0000

```
LR test: independent vs. saturated:  chi2(66) = 4320.87 Prob>chi2 = 0.0000
```

```
Factor loadings (pattern matrix) and unique variances
```

Table. The result of factor analysis

Based on the results of the analysis, 4 factors were selected, since their own values are greater than 1.

```
LR test: independent vs. saturated:  chi2(66) = 4320.87 Prob>chi2 = 0.0000
```

```
Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
J721637	0.3358	-0.0192	-0.3451	0.5407	0.4754
J721631	0.5401	-0.0529	0.1987	-0.3222	0.5622
J721632	0.5877	-0.1112	-0.3047	0.1161	0.5359
J721633	0.7434	0.0609	0.1843	-0.2168	0.3626
J721634	0.7749	0.1243	0.2356	-0.1229	0.3134
J721636	0.7883	0.0364	0.1258	-0.0662	0.3571
J721635	0.7274	-0.0385	-0.0466	0.0468	0.4651
L5	-0.0445	0.2668	0.6641	0.5598	0.1724
M80	0.1123	0.7713	-0.1736	0.1430	0.3420
M71	-0.1983	0.6686	-0.1719	-0.3444	0.3655
J6	0.3458	0.0298	0.0169	0.2738	0.8043
J65_new	-0.4302	-0.0131	0.4371	-0.1038	0.6129

Table. Correlation matrix.

```

. rotate, blanks (0.7)

Factor analysis/correlation                                Number of obs    =    1,973
  Method: principal-component factors                      Retained factors  =         4
  Rotation: orthogonal varimax (Kaiser off)                Number of params  =        42

```

Factor analysis/correlation	Number of obs	=	1,973
Method: principal-component factors	Retained factors	=	4
Rotation: orthogonal varimax (Kaiser off)	Number of params	=	42

LR test: independent vs. saturated: $\chi^2(66) = 4320.87$ Prob> $\chi^2 = 0.0000$

We rotated the model to get the best result. As a result of the rotation, 4 factors were obtained, of which Factor 1 describes 25% of the model, Factor 2 - 12%, Factor 3 – 10%, factor 4 – 9%. Together, the factors describe 55.2% of the model. Of course, this result is not perfect, but it is already considered very good.

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
J721637		0.7186			0.4754
J721631					0.5622
J721632					0.5359
J721633	0.7969				0.3626
J721634	0.8187				0.3134
J721636	0.7771				0.3571
J721635					0.4651
L5				0.9093	0.1724
M80			0.7616		0.3420
M71			0.7343		0.3655
J6					0.8043
J65_new					0.6129

Factor rotation matrix

We set the limit of "getting the variable into the factor" at 0.7. Based on the results of this boundary, we can say that:

Factor 1 "consists" of variable answers to the questions "Do you or your family have the opportunity, if you wish, to save money for large purchases - a car, a cottage?", "Do you or your family have the opportunity, if you wish, to spend a vacation abroad with the whole family ?", "Do you or your family have the opportunity , if you wish, to spend a vacation with the whole family on a Russian resort?"

Factor 2 – from the variable of the answer to the question "Do you or your family have the opportunity to eat meat, chicken or fish every other day?"

Factor 3 – from the variables of the answer to the questions "Do you currently smoke?", "Have you consumed alcoholic beverages in the last 30 days?"

Factor 4 – from the variable of the answer to the question "Have you had any health problems in the last 30 days?".

We believe that the result turned out to be very interesting, since initially we combined these variables in meaning for ourselves in a slightly different way - the variables from factor 3 for us were variables of the same semantic block (they turned out to be in the same block according to the results), but we did not see for ourselves the divisions between questions about fish, house, apartment, resorts.

As a result, factor 1 can be conditionally called the factor of the possibility of making large material purchases, factor 2 - the possibility of eating meat products, factor 3 - the factor of bad habits, factor 4 - the health factor.

```
. estat kmo, novar
```

```
Kaiser-Meyer-Olkin measure of sampling adequacy (overall) = 0.8386
```

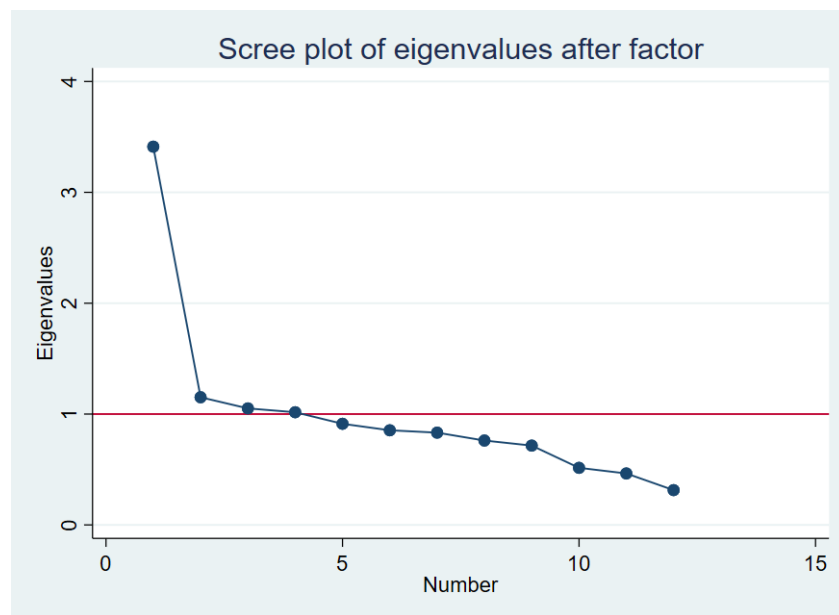
The KMO test verifies the applicability of the resulting model to the "realities of the database". In our case , it tells us that we got a very good result (0.838) – "high adequacy of the model" (from 0.8 to 0.9). We did not expect that the estimate of our factor model would be so high.

```
. predict factor1 factor2 factor3 factor4
(regression scoring assumed)
```

Scoring coefficients (method = regression; based on varimax rotated factors)

Variable	Factor1	Factor2	Factor3	Factor4
J721631	0.28436	-0.26792	-0.05816	-0.08579
J721637	-0.15484	0.60700	-0.00234	0.09120
J721632	0.03506	0.32212	-0.03937	-0.17192
J721633	0.31294	-0.16704	0.02509	-0.00612
J721634	0.31344	-0.12855	0.05367	0.10220
J721636	0.26584	-0.02365	-0.00014	0.04135
J721635	0.16656	0.14566	-0.03408	-0.02303
L5	0.03146	0.00974	-0.00610	0.86909
M80	-0.00789	0.17759	0.66526	0.15006
M71	0.03758	-0.19576	0.64221	-0.17265
J6	0.02338	0.22130	-0.01632	0.18430
J65_new	0.03195	-0.36274	-0.09921	0.23874

Let's estimate the degree of correlation of our variables with the resulting factors. In principle, the correlation values that we obtained are adequate. A sufficient number of medium and strong connections.



Construction and estimation of binary regression

Since initially our variable denoting the level of life satisfaction is ranked, but satisfaction with life as a whole can be defined as "a person is satisfied with it" and "a person is not satisfied with it ", we recoded it into a binary one and decided Build a binary regression based on ordinal variables.

logit J65_newnew J1 _v8 _v9 _v10 _v11 _v451 M3 J63 J64

```
. logit J65_newnew J1 _v8 _v9 _v10 _v11 _v451 M3 J63 J64
```

```
Iteration 0:  log likelihood = -4856.8809
Iteration 1:  log likelihood = -4280.3636
Iteration 2:  log likelihood = -4278.1541
Iteration 3:  log likelihood = -4278.1529
Iteration 4:  log likelihood = -4278.1529
```

```
Logistic regression               Number of obs   =      7,090
                                LR chi2(9)         =     1157.46
                                Prob > chi2         =      0.0000
Log likelihood = -4278.1529       Pseudo R2        =      0.1192
```

J65_newnew	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
J1	.6516195	.1513277	4.31	0.000	.3550226	.9482163
_v8	-.2827047	.048411	-5.84	0.000	-.3775884	-.1878209
_v9	-.0940785	.0447723	-2.10	0.036	-.1818305	-.0063264
_v10	-.2263101	.030079	-7.52	0.000	-.2852639	-.1673563
_v11	-.1123488	.0325618	-3.45	0.001	-.1761688	-.0485288
_v451	-.1095306	.0284415	-3.85	0.000	-.165275	-.0537862
M3	-.5383808	.0455876	-11.81	0.000	-.6277308	-.4490307
J63	.1528627	.0176918	8.64	0.000	.1181875	.187538
J64	.1287891	.0160375	8.03	0.000	.097356	.1602221
_cons	1.640984	.2536775	6.47	0.000	1.143785	2.138183

Table. Model with coefficients

```
. logistic J65_newnew J1 _v8 _v9 _v10 _v11 _v451 M3 J63 J64
```

Logistic regression	Number of obs	=	7,090
	LR chi2(9)	=	1157.46
	Prob > chi2	=	0.0000
Log likelihood = -4278.1529	Pseudo R2	=	0.1192

J65_newnew	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
J1	1.918646	.2903442	4.31	0.000	1.426213	2.581102
_v8	.7537424	.0364894	-5.84	0.000	.6855126	.8287631
_v9	.9102113	.0407522	-2.10	0.036	.8337426	.9936935
_v10	.7974708	.0239871	-7.52	0.000	.7518158	.8458982
_v11	.8937325	.0291016	-3.45	0.001	.8384764	.9526299
_v451	.8962547	.0254909	-3.85	0.000	.8476606	.9476347
M3	.5836926	.0266091	-11.81	0.000	.5338017	.6382465
J63	1.165165	.0206138	8.64	0.000	1.125455	1.206276
J64	1.13745	.0182419	8.03	0.000	1.102253	1.173772
_cons	5.160245	1.309038	6.47	0.000	3.138627	8.484006

Note: `_cons` estimates baseline odds.

Table. A model with odds ratios

The odds ratio values are greater than 1 for the following significant variables: J1, J63, J64. So, we can say that

- an increase in the value of J1 (different types of employment (from more "busy" to "less")) by 1 will lead to an increase in the chance of being satisfied with one's life by 1.91 times;
- an increase in the value of J63 (means self-assessment of the level of possession of power) by 1 will lead to an increase in the chance of being satisfied with one's life by 1.165 times;
- an increase in the value of J64 (means self-assessment of the level of having respect in society) by 1 will lead to an increase in the chance of being satisfied with one's life by 1,137 times.

Let us evaluate the significance and predictive power of the regression model.

`Prob > chi2` = **0.0000**

This suggests that the model is significant and we can accept our H0 hypothesis that there are no differences between the values predicted by our model and those actually observed.

```
. estat classification
```

```
Logistic model for J65_newnew
```

Classified	True		Total
	D	~D	
+	3077	1431	4508
-	919	1663	2582
Total	3996	3094	7090

```
Classified + if predicted Pr(D) >= .5
```

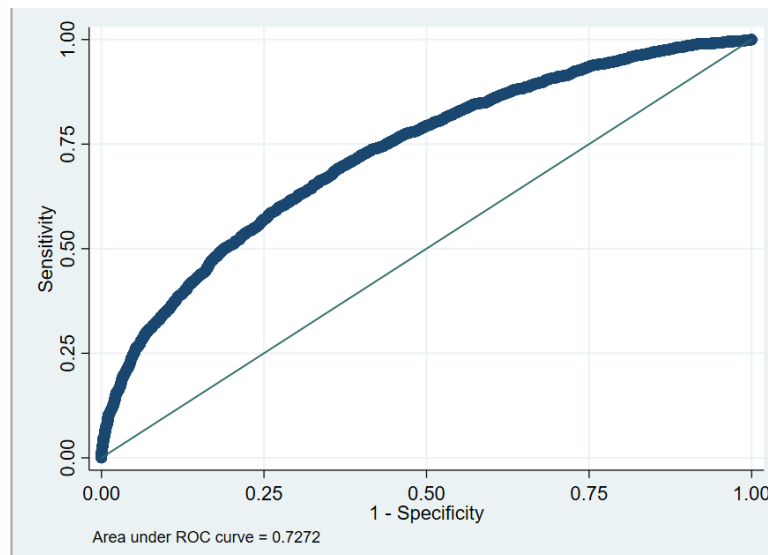
```
True D defined as J65_newnew != 0
```

Sensitivity	Pr(+ D)	77.00%
Specificity	Pr(- ~D)	53.75%
Positive predictive value	Pr(D +)	68.26%
Negative predictive value	Pr(~D -)	64.41%
False + rate for true ~D	Pr(+ ~D)	46.25%
False - rate for true D	Pr(- D)	23.00%
False + rate for classified +	Pr(~D +)	31.74%
False - rate for classified -	Pr(D -)	35.59%
Correctly classified		66.85%

67% of cases are correctly predicted. This is a pretty good result. The constructed regression model is statistically significant, since the coefficient is 0.000, which is less than 0.05. This means that this model describes the situation better than the use of a more primitive model of the mean.

ROC curve graph ROC curve (receiver operating characteristic) is a graph that allows you to evaluate the quality of a binary classification. Also known as the error curve.

The quantitative interpretation of ROC is given by the AUC indicator (area under ROC curve, area under the ROC curve) - the area bounded by the ROC curve and the axis of dols and false positive classifications. The higher the AUC, the better the classifier, while the value of 0.5 demonstrates the unsuitability of the chosen classification method (corresponds to random divination).



Our indicator is 0.72. We consider this a good result, since the model is unambiguous
 "Not just guessing."

Let's write the regression equation.

The probability of an event occurring is calculated using the formula $p = \frac{1}{1 + e^{-z}}$
 $z = b_1x_1 + b_2x_2 + \dots + b_nx_n$

x_n are the values of the explanatory variables,

b_n - coefficients, the calculation of which is a problem of binary logistic regression,

a is a constant.

In our case, the regression equation is:

$$J65_newnew = 1,64 - 0,65 * J1 - 0,28 * v8 - 0,9 * v9 - 0,22 * v10 - 0,1 * v11 - 0,1 * v451 - 0,5 * M3 + 0,15 * J63 + 0,12 * J64$$

Note that the influence of each of the independent variables on the dependent variable is statistically significant.

We kept the probabilities and balances. Let's assess the normality. Normality Assessment:

Hypothesis H0: the distribution of variable values does not differ significantly from the normal distribution.

Hypothesis H1: The distribution of variable values is significantly different from the normal distribution.

Let's use the Kolmogorov-Smirnov test, because. The sample is greater than 100.

```
egen r_mu = mean(r)
```

```
egen r_s = sd(r)
```

```
ksmirnov r = normprob((r-r_mu)/r_s)
```

```
. ksmirnov r = normprob((r-r_mu)/r_s)
```

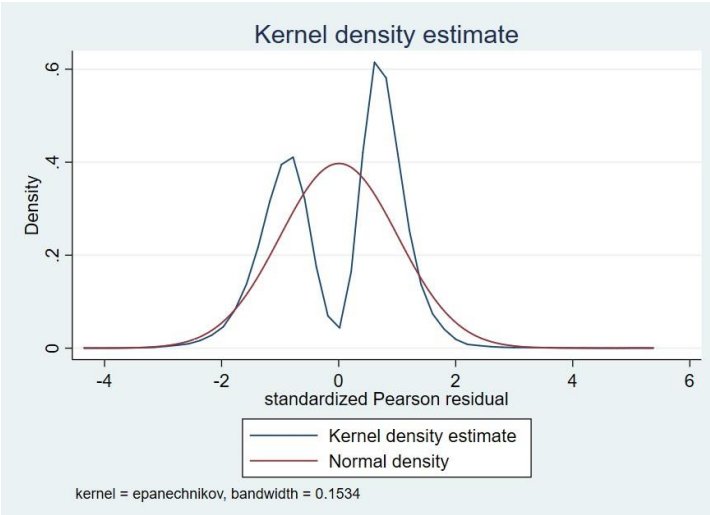
One-sample Kolmogorov-Smirnov test against theoretical distribution
normprob((r-r_mu)/r_s)

Smaller group	D	P-value
r:	0.0979	0.000
Cumulative:	-0.1601	0.000
Combined K-S:	0.1601	0.000

As a result, a significance of less than 0.05 was obtained, therefore, this distribution is significantly different from the normal one.

Graph

kdensity r, normal



The graph confirms the normality test performed.

Let's identify outliers or influencing observations.

An outlier is an observation that doesn't fit into the model. We look for emissions by residues. Standardized residues should be from -3 to +3 (3-sigma rule). Look at those residues that modulo 3, these are potential outliers.

```
tabulate r if r > 3
```

```
. tabulate r if r > 3
```

standardize d Pearson residual	Freq.	Percent	Cum.
3.023821	1	8.33	8.33
3.029408	1	8.33	16.67
3.149669	1	8.33	25.00
3.32839	1	8.33	33.33
3.439364	1	8.33	41.67
3.487803	1	8.33	50.00
3.525684	1	8.33	58.33
3.794357	1	8.33	66.67
4.004742	1	8.33	75.00
4.36672	1	8.33	83.33
5.00767	1	8.33	91.67
5.226831	1	8.33	100.00
Total	12	100.00	

tabulate r if r < -3

```
. tabulate r if r < -3
```

standardize d Pearson residual	Freq.	Percent	Cum.
-4.201677	1	14.29	14.29
-3.499016	1	14.29	28.57
-3.423597	1	14.29	42.86
-3.185069	1	14.29	57.14
-3.061734	1	14.29	71.43
-3.040442	1	14.29	85.71
-3.011945	1	14.29	100.00
Total	7	100.00	

There are 19 outliers in our model. However , if we consider that we are considering several thousand values, then this is not critical.

Let us estimate the multicollinearity using the correlation matrix pwcorr

J65_newnew J1 _v8 _v9 _v10 _v11 _v451 J10 M3 J63 J64 , sig

. pwcorr J65_newnew J1 _v8 _v9 _v10 _v11 _v451 J10 M3 J63 J64, sig								
	J65~wnew	J1	_v8	_v9	_v10	_v11	_v451	
J65_newnew	1.0000							
J1	-0.1139 0.0000	1.0000						
_v8	-0.2685 0.0000	0.0364 0.0014	1.0000					
_v9	-0.2465 0.0000	0.0141 0.2135	0.7383 0.0000	1.0000				
_v10	-0.2735 0.0000	0.0073 0.5229	0.5443 0.0000	0.5172 0.0000	1.0000			
_v11	-0.2420 0.0000	0.0276 0.0180	0.5871 0.0000	0.5554 0.0000	0.5891 0.0000	1.0000		
_v451	-0.0461 0.0000	-0.0629 0.0000	0.0347 0.0027	0.0222 0.0553	0.0274 0.0181	0.0450 0.0001	1.0000	
J10	0.1620 0.0000	-0.6452 0.0000	-0.1464 0.0000	-0.1092 0.0000	-0.2378 0.0000	-0.1215 0.0000	0.0474 0.0000	
M3	-0.2430 0.0000	0.2861 0.0000	0.1356 0.0000	0.1292 0.0000	0.1477 0.0000	0.1126 0.0000	-0.0680 0.0000	
J63	0.2670 0.0000	-0.1086 0.0000	-0.1575 0.0000	-0.1450 0.0000	-0.2033 0.0000	-0.1521 0.0000	-0.0416 0.0000	
J64	0.2184	-0.1041	-0.1745	-0.1543	-0.1333	-0.1531	-0.0538	

	J10	M3	J63	J64
J10	1.0000			
M3	-0.2261 0.0000	1.0000		
J63	0.1217 0.0000	-0.2485 0.0000	1.0000	
J64	0.1288 0.0000	-0.1104 0.0000	0.4259 0.0000	1.0000

We see that the correlation coefficients in all cases except eight are less than 0.5, taking into account the number of variables, we can say that the correlation of variables with each other is mostly weak.

However, we cannot claim that we do not have multicollinearity, since the correlation coefficient $_v8$ and $_v9 = 0.738$, $J10$ and $J1 = -0.64$, etc.

Construction and estimation of linear regression

We decided to do linear regression as part of the study to demonstrate our proficiency in this tool. We were faced with a limitation related to the peculiarity of the types of our variables, so it was decided to assess the influence of factors on salary, since we had previously established that salary, in principle, has an impact on the level of life satisfaction.

Variable recoding

Linear regression involves the use of interval or binary variables as predictors and an interval variable as a dependent variable. Also, in order to build a binary regression model, it is necessary to use a binary variable as a dependent variable, and either interval or binary variables as predictors. However, in our initial database, there are only 4 interval variables (age, work experience, salary, number of vacation days) and There are a lot of ordinal variables (level of education, availability of children, assessment of health problems, availability of subordinates, etc.).

In order to use the potential of these ordinal variables, as well as to make up for the insufficient number of interval variables to build a variety of regression models, we decided to create new binary variables based on ordinal variables.

Commands used to convert ordinal variables to binary variables:

recode DIPLOM (1=0) (2=0) (3=0) (4=0) (5=1) (6=1), gen (DIPLOM_new)

recode subordinate (1=1) (2=0), gen (subordinate_new)

recode health_problems (1=0) (2=1), gen (health_problems_new)

Variables denoting age and work experience are excluded from the analysis in order to reduce the multicollinearity of the model.

Regression:

reg salary DIPLOM_new children health_problems_new vacation power respect

subordinate_new

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
DIPLOM_new	3189.944	504.6226	6.32	0.000	2200.624	4179.264
children	1899.933	581.4627	3.27	0.001	759.9672	3039.899
health_problems_new	1544.017	530.2606	2.91	0.004	504.4334	2583.6
vacation	29.29837	15.85955	1.85	0.065	-1.794509	60.39124
power	344.6223	158.6859	2.17	0.030	33.51627	655.7284
respect	730.4003	166.1129	4.40	0.000	404.7337	1056.067
subordinate_new	8409.866	567.5698	14.82	0.000	7297.138	9522.595
_cons	13717.46	1399.954	9.80	0.000	10972.83	16462.09

The number of vacation days is not a significant variable, so we do not include it in the equation.

The regression equation is:

$$\begin{aligned} \text{salary} = & 3189.944 \cdot \text{DIPLOM_new} + 1899.933 \cdot \text{children} + 1544.017 \\ & \cdot \text{health_problems_new} + 344.6223 \cdot \text{power} + 730.4 \cdot \text{respect} + 8409.866 \\ & \cdot \text{subordinate_new} \end{aligned}$$

Let's study the standardized regression coefficients:

reg salary DIPLOM_new children health_problems_new vacation power respect
subordinate_new, beta

salary	Coef.	Std. Err.	t	P> t	Beta
DIPLOM_new	3189.944	504.6226	6.32	0.000	.0949477
children	1899.933	581.4627	3.27	0.001	.0480602
health_problems_new	1544.017	530.2606	2.91	0.004	.042589
vacation	29.29837	15.85955	1.85	0.065	.0273383
power	344.6223	158.6859	2.17	0.030	.0347726
respect	730.4003	166.1129	4.40	0.000	.0705701
subordinate_new	8409.866	567.5698	14.82	0.000	.2215156
_cons	13717.46	1399.954	9.80	0.000	.

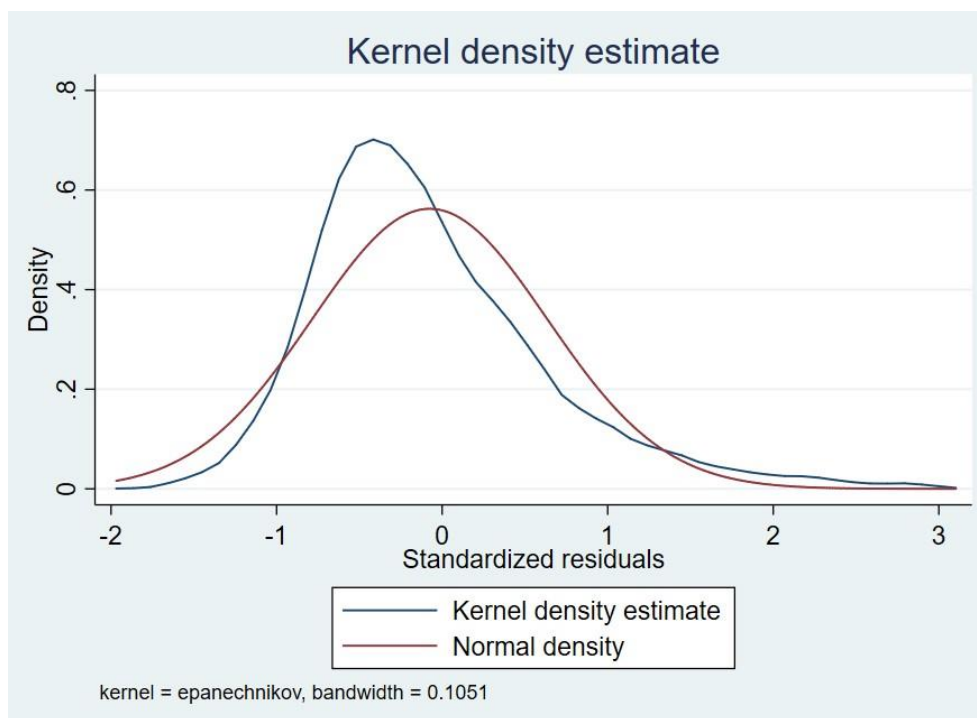
As we can see, the subordinate_new has the greatest relationship with the salary variable , i.e. the change in salary is greatly affected by the presence of subordinates in a person. Next in terms of the strength of the connection with the salary is a variable showing the level of education. Respect for the person also plays a role.

Diagnosing the model

First, we standardize the leftovers using the command
predict r, rstandard

Then we check the normality of the residues using
swilk r

The residues are abnormally distributed. Graph:
kdensity r, normal



Outliers and influencing observations

Outliers are measurement results that stand out from the total sample. If the standardized residue (r) modulo is greater than 3 (the 3-sigma rule), then the value is an outlier.

We sort the residues and notice that we get 62 outliers

Total	62	100.00
-------	----	--------

As for the influencing observations, the calculation of the influence statistics is made using the dfbeta command. We have no influencing observations

Multicollinearity test:

To check multicollinearity, we use the values of the VIF indicator and the correlation matrix. The estat vif command allows you to find out the strength of the correlation of a variable with other explanatory variables.

estat vif

```
. estat vif
```

Variable	VIF	1/VIF
respect	1.21	0.823416
power	1.21	0.827333
DIPLOM_new	1.06	0.940175
subordinat~w	1.05	0.949022
vacation	1.03	0.968520
children	1.02	0.980408
health_pro~w	1.01	0.991463
Mean VIF	1.09	

Multicollinearity is considered to exist if the total VIF value is >5. If it is at the level of 1-5, then there is no multicollinearity.

Thus, we built a linear regression, which helped us understand what affects the level of wages.

Conclusion

Our research allowed us to learn many interesting facts about the level of life satisfaction of the inhabitants of Russia. Let us dwell on the main hypotheses, the verification of which we proposed to observe in the course of acquaintance with the results of our study.

1) Women feel more satisfied with life than men – in the course of checking with the help of a comparison of averages, it turned out that we cannot say this, rather, men are more satisfied;

2) Residents of cities are satisfied with life more than residents of villages and towns - we can not say that there is a clear relationship between the type of settlement and the level of satisfaction;

3) There is a significant relationship between the state of health and the level of life satisfaction
- Yes, indeed, according to the variable that stores data on self-assessment of health, we were able to understand that there is a direct significant relationship. Note, however, that it is more likely to be estimated closer to the average than to the strong.

4) The "degree" of religiosity can significantly affect the level of satisfaction with life, and believers are more satisfied with it - no, this is not so.

5) People who feel respected, feel satisfied with life – yes, that's true, and the regression model has proven it.

6) People who think they have a lot of power feel satisfied with life – this is also true, which was also found out during the construction of the regression model.

As part of the study, we were also able to draw many interesting conclusions. For example, a high salary by the standards of society is not a key factor in the formation of the level of life satisfaction.

Sources of information

- 1) Joint database of 1994-2018 of the results of the ODA of individual individuals of the RLMS of the Higher School of Economics. URL: <https://www.hse.ru/rlms/spss>. Access mode: free (accessed 16.05.2020)
- 2) Nastina E.A. and Almakaeva A.M. The role of the level of claims and social comparisons in determining life satisfaction. URL: <https://cyberleninka.ru/article/n/rol-urovnya-prityazaniy-i-sotsialnyh-sravneniy-v-determinatsii-udovletvorennosti-zhiznyu> Access mode: free (accessed 16.05.2020)
- 3) Balatsky E.V. Life Satisfaction Factors : Measurement and Integral Indicators. URL: <https://cyberleninka.ru/article/n/factory-udovletvorennosti-zhiznyu-izmerenie-i-integralnye-pokazateli/viewer>, access mode: free (accessed 16.05.2020)
- 4) Rosstat called the ratio of the number of men and women in Russia. URL: <https://ria.ru/20200422/1570386332.html>, access mode: free (accessed 17.05.2020)
- 5) Urbanization in Russia. Rosinfostat. URL: <https://rosinfostat.ru/urbanizatsiya-v-rossii/>, access mode: free (accessed 18.05.2020)