

Задание

По данным, представленным в таблице 1, изучаются коэффициенты загрузки сотовых сетей в разных странах Y от переменных:

- X_1 (Количество вышек сотовой связи)
- X_2 (Среднее количество звонков утром относительно 1 вышки)
- X_3 (Среднее количество звонков днем относительно 1 вышки)
- X_4 (Среднее количество звонков вечером относительно 1 вышки)
- X_5 (Количество абонентов сотовой связи, проживающих в зоне покрытия 1 вышки)
- X_6 (Среднее количество звонков ночью относительно 1 вышки)

1. Построить уравнение множественной регрессии в линейной форме с полным набором факторов.

2. Построить матрицу парных коэффициентов корреляции. Установить какие факторы мультиколлинеарные.

3. Оценить статистическую зависимость уравнения регрессии и его параметров с помощью t -критерия Стьюдента и F -критерия Фишера.

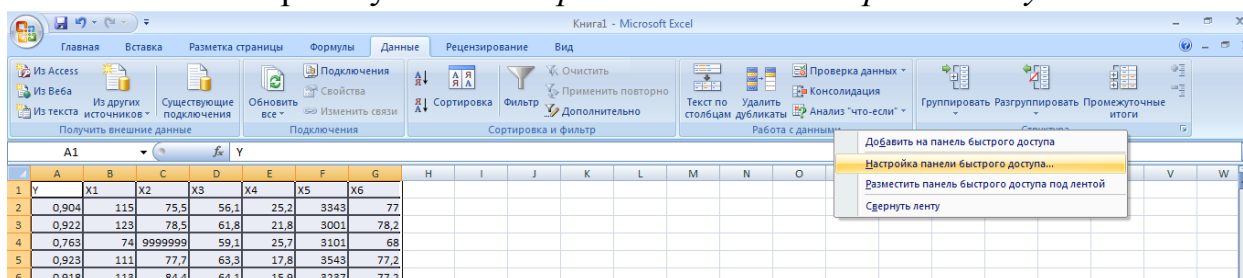
4. Отобрать информативные факторы. Построить уравнение регрессии со статистически значимыми факторами. Рассчитать среднюю ошибку аппроксимации для полученной модели.

Таблица 1.

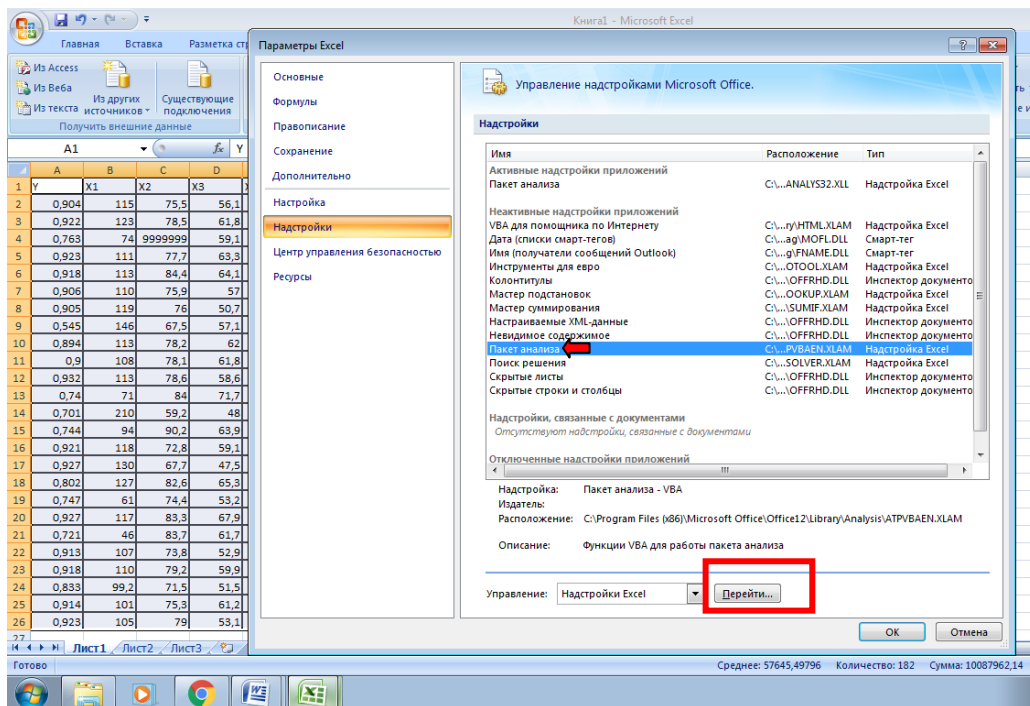
Страна	Y	X1	X2	X3	X4	X5	X6
Австрия	0,904	115	75,5	56,1	25,2	3343	77
Австралия	0,922	123	78,5	61,8	21,8	3001	78,2
Белоруссия	0,763	74	78,4	59,1	25,7	3101	68
Бельгия	0,923	111	77,7	63,3	17,8	3543	77,2
Великобритания	0,918	113	84,4	64,1	15,9	3237	77,2
Германия	0,906	110	75,9	57	22,4	3330	77,2
Дания	0,905	119	76	50,7	20,6	3808	75,7
Индия	0,545	146	67,5	57,1	25,2	2415	62,6
Испания	0,894	113	78,2	62	20,7	3295	78
Италия	0,9	108	78,1	61,8	17,5	3504	78,2
Канада	0,932	113	78,6	58,6	19,7	3056	79
Казахстан	0,74	71	84	71,7	18,5	3007	67,6
Китай	0,701	210	59,2	48	42,4	2844	69,8
Латвия	0,744	94	90,2	63,9	23	2861	68,4
Нидерланды	0,921	118	72,8	59,1	20,2	3259	77,9
Норвегия	0,927	130	67,7	47,5	25,2	3350	78,1
Польша	0,802	127	82,6	65,3	22,4	3344	72,5
Россия	0,747	61	74,4	53,2	22,7	2704	66,6
США	0,927	117	83,3	67,9	18,1	3642	76,7
Украина	0,721	46	83,7	61,7	20,1	2753	68,8
Финляндия	0,913	107	73,8	52,9	17,3	2916	76,8
Франция	0,918	110	79,2	59,9	16,8	3551	78,1
Чехия	0,833	99	71,5	51,5	29,9	3177	73,9
Швейцария	0,914	101	75,3	61,2	20,3	3280	78,6
Швеция	0,923	105	79	53,1	14,1	3160	78,5

1. Построим линейную регрессионную модель по всему массиву.

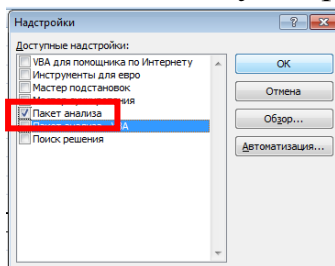
Для этого воспользуемся “пакетом анализа”, расположенным на вкладке “Данные”. Если данный пакет не активирован в вашей версии Excel, необходимо добавить его на панель быстрого доступа вручную. Для этого на вкладке “Данные” щелкаем правой кнопкой мыши на блок «Работа с данными» и выберем пункт «Настройка панели быстрого доступа»



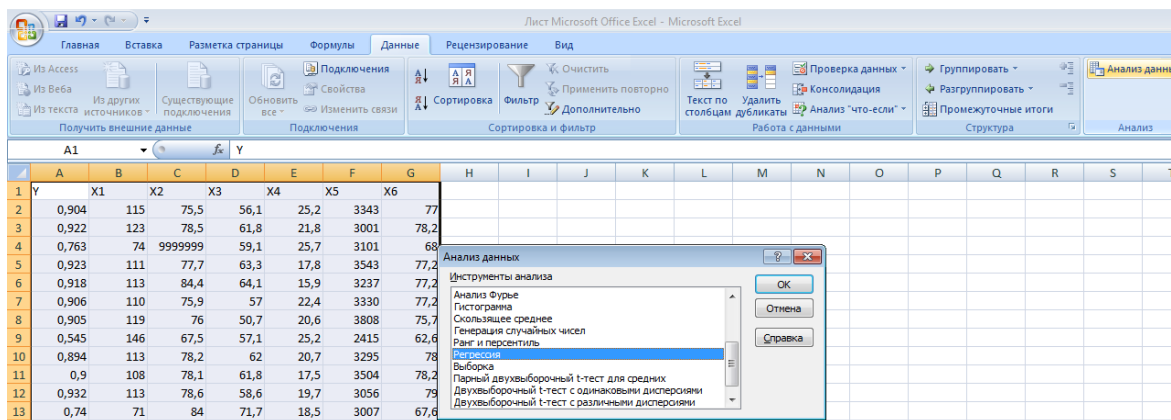
В появившемся окне выбираем пункт «Надстройки», → «Пакет анализа» → нажимаем кнопку «Перейти»



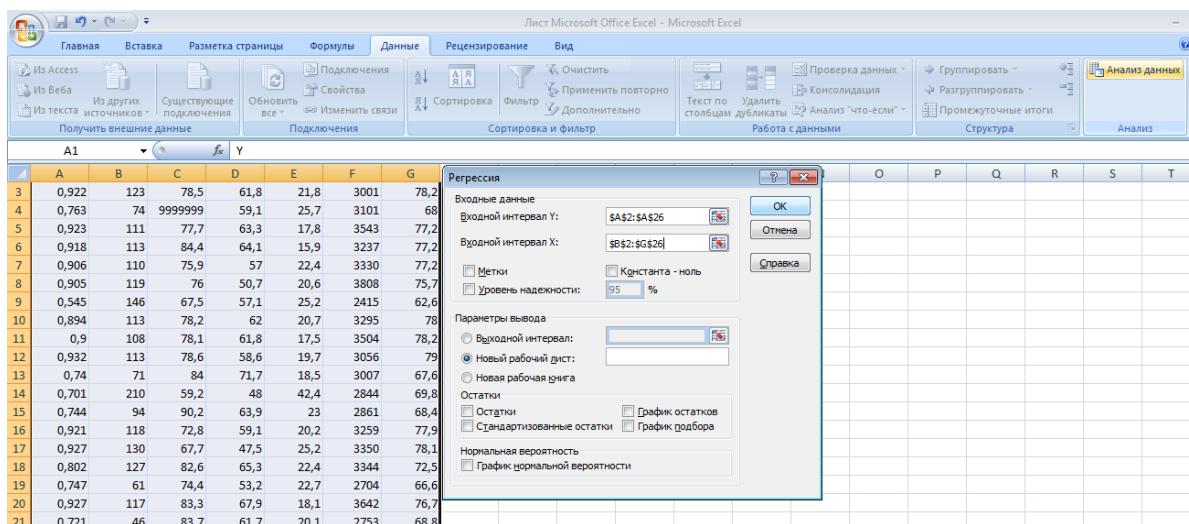
Ставим галочку напротив пункта «Пакет анализа» и нажимаем «ОК»



В правом верхнем углу на вкладке «Данные» появится блок «Анализ данных». Для того чтобы построить регрессионную модель с полным набором факторов выбираем в блоке «Анализ данных» элемент «регрессия» и задаем входные интервалы для анализа.



В качестве входного интервала Y выбираем все значения столбца Y, качестве входного интервала X выбираем значения всех столбцов переменной X (6 столбцов).



По умолчанию результаты будут выведены на новый лист. Проанализируем полученные данные.

Таблица 2.

Регрессионная статистика	
Множественный R	0,981982
R-квадрат	0,964288
Нормированный R-квадрат	0,952384
Стандартная ошибка	0,022104
Наблюдения	25

Величина R-квадрат, называемая также мерой определенности, характеризует качество полученной регрессионной прямой. Это качество выражается степенью соответствия между исходными данными и регрессионной моделью (расчетными данными). Мера определенности всегда находится в пределах интервала $[0;1]$.

Если значение R-квадрата близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение R-квадрата, близкое к нулю, означает плохое качество построенной модели.

Полученный результат, равный 0,964288, говорит о хорошей подгонке регрессионной прямой к исходным данным.

Таблица 3.

Дисперсионный анализ					
	df	SS	MS	F	Значимость F
Регрессия	6	0,237468	0,039578	81,00515	4,87E-12

Остаток	18	0,008795	0,000489		
Итого	24	0,246263			

Таблица 4.

	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-значения</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>	<i>Нижние 95,0%</i>	<i>Верхние 95,0%</i>
У-пересечение	-0,63627	0,154851	-4,10892	0,000659	-0,9616	-0,31094	-0,9616	-0,31094
Переменная X 1	-0,00048	0,000228	-2,08803	0,051278	-0,00095	2,94E-06	-0,00095	2,94E-06
Переменная X 2	0,001308	0,001462	0,894695	0,382758	-0,00176	0,004381	-0,00176	0,004381
Переменная X 3	-0,00141	0,001235	-1,1414	0,268661	-0,004	0,001185	-0,004	0,001185
Переменная X 4	0,000149	0,001392	0,106848	0,916091	-0,00277	0,003072	-0,00277	0,003072
Переменная X 5	4,32E-05	2,04E-05	2,114043	0,048729	2,68E-07	8,62E-05	2,68E-07	8,62E-05
Переменная X 6	0,018535	0,00159	11,65416	8,07E-10	0,015194	0,021877	0,015194	0,021877

Исходя из расчетов, можем записать уравнение регрессии таким образом:

$$y = -0,632627 - 0,00048 * x_1 + 0,001308 * x_2 - 0,00141 * x_3 + 0,000149 * x_4 + 4,32E-05 * x_5 + 0,018535 * x_6$$

Используя пакет анализа MS Excel и инструмент анализа «Корреляция», можем построить корреляционную матрицу. Выбираем в блоке «Анализ данных» элемент «Корреляция» и задаем входные интервалы для анализа. В качестве входного интервала задаем всю таблицу. Результаты также, по умолчанию, будут выведены на новый лист.

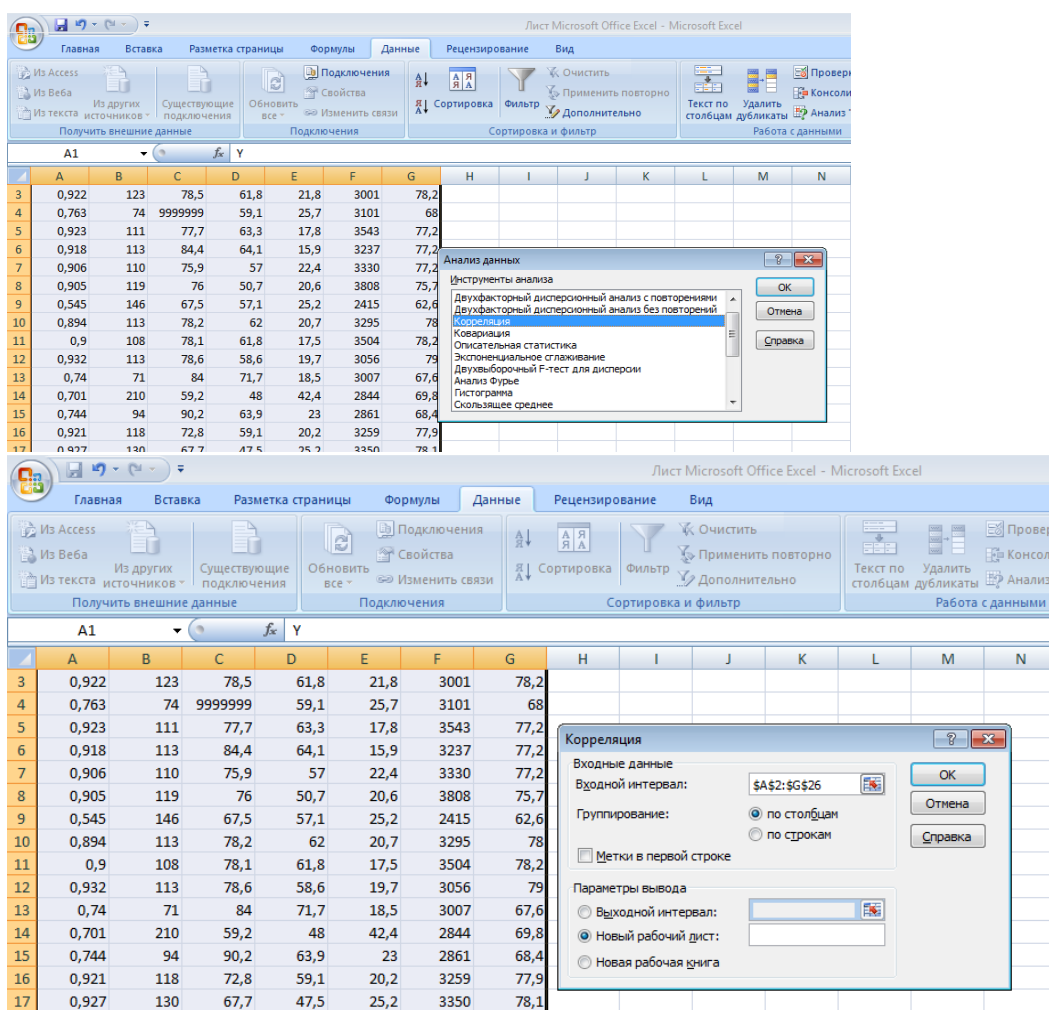


Таблица 5.

	Y	X1	X2	X3	X4	X5	X6
Y	1						
X1	-0,00434	1					
X2	0,17050	-0,62897	1				
X3	-0,00433	-0,36511	0,76495	1			
X4	-0,48711	0,54107	-0,66713	-0,49626	1		
X5	0,75145	0,07786	0,18551	0,10997	-0,33127	1	
X6	0,96203	0,16328	0,04856	-0,05212	-0,40689	0,70393	1

Анализ таблицы 5 показывает наличие линейной связи между признакам X1, X2, X4 и X6.

Исключив мультиколлиниарные факторы, еще раз построим регрессионную модель и оценим статистическую зависимость уравнения регрессии.

Выбираем в блоке «Анализ данных» элемент «Регрессия» и задаем входные интервалы для анализа.

Таблица 6.

Страна	Y	X1	X2	X4	X6
Австрия	0,904	115	75,5	25,2	77
Австралия	0,922	123	78,5	21,8	78,2
Белоруссия	0,763	74	78,4	25,7	68
Бельгия	0,923	111	77,7	17,8	77,2
Великобритания	0,918	113	84,4	15,9	77,2
Германия	0,906	110	75,9	22,4	77,2
Дания	0,905	119	76	20,6	75,7
Индия	0,545	146	67,5	25,2	62,6
Испания	0,894	113	78,2	20,7	78
Италия	0,9	108	78,1	17,5	78,2
Канада	0,932	113	78,6	19,7	79
Казахстан	0,74	71	84	18,5	67,6
Китай	0,701	210	59,2	42,4	69,8
Латвия	0,744	94	90,2	23	68,4
Нидерланды	0,921	118	72,8	20,2	77,9
Норвегия	0,927	130	67,7	25,2	78,1
Польша	0,802	127	82,6	22,4	72,5
Россия	0,747	61	74,4	22,7	66,6
США	0,927	117	83,3	18,1	76,7
Украина	0,721	46	83,7	20,1	68,8
Финляндия	0,913	107	73,8	17,3	76,8
Франция	0,918	110	79,2	16,8	78,1
Чехия	0,833	99	71,5	29,9	73,9
Швейцария	0,914	101	75,3	20,3	78,6
Швеция	0,923	105	79	14,1	78,5

Таблица 7.

<i>Регрессионная статистика</i>	
Множественный R	0,976382
R-квадрат	0,953322
Нормированный R-квадрат	0,943987
Стандартная ошибка	0,023974
Наблюдения	25

Таблица 8.

Дисперсионный анализ					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
Регрессия	4	0,237468	0,058692	102,1176	5,17E-13
Остаток	20	0,011495	0,000575		
Итого	24	0,246863			

Таблица 9.

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	-0,72682	0,16265	-4,4686	0,000235	-1,0661	-0,38754	-1,0661	-0,38754
Переменная X 1	-0,00053	0,000233	-2,25681	0,035366	-0,00101	-4E-05	-0,00101	-4E-05
Переменная X 2	0,000749	0,00115	0,650989	0,522463	-0,00165	0,003149	-0,00165	0,003149
Переменная X 4	0,000669	0,001479	0,4524	0,65585	-0,00242	0,003753	-0,00242	0,003753
Переменная X 6	0,020988	0,001294	16,22046	5,64E-13	0,018289	0,023687	0,018289	0,023687

Обратите внимание – нумерация переменных X происходит автоматически, средствами MS Excel. Необходимо установить соответствие переменных вручную (выделено в таблице цветом).

Согласно таблице 9, можем записать уравнение регрессии таким образом:

$$y = -0,72682 - 0,00053 * x_1 + 0,000749 * x_2 + 0,000669 * x_4 + 0,020988 * x_6$$

Исходя из условия, что если ноль попадает доверительный интервал, то фактор является незначимым, проведем анализ последних двух колонок таблицы 11.

Таблица 11.

	Коэффициенты	Нижние 95,0%	Верхние 95,0%
Y-пересечение	-0,72682	-1,0661	-0,38754
Переменная X 1	-0,00053	-0,00101	-4E-05
Переменная X 2	0,000749	-0,00165	0,003149

Переменная X 4	0,000669	-0,00242	0,003753
Переменная X 6	0,020988	0,018289	0,023687

Анализ показывает, что для X1 и X6 «ноль» не попадает в доверительный интервал, следовательно, факторы являются значимыми.

Значимость уравнения регрессии оценивается с помощью F-критерия Фишера. Для этого воспользуемся результатами из таблицы 12.

Таблица 12.

	Линейная
Значимость F	102,1176

Для определения F_T определим число степеней свободы

$$(25(\text{число наблюдений}) - 4(\text{число факторов в модели}) - 1 = 20)$$

и воспользуемся вспомогательной таблицей. Таким образом, сравнивая табличное значение ($F_T = 2,87$) с расчетным выполняется неравенство $F_p > F_T$, следовательно делаем вывод, что коэффициент чистой регрессии b_i при факторе x_i статистическим значим.

Построим уравнение регрессии на основе информативных факторов.

Таблица 13.

Регрессионная статистика	
Множественный R	0,975847
R-квадрат	0,952277
Нормированный R-квадрат	0,947938
Стандартная ошибка	0,023113
Наблюдения	25

Таблица 14.

Дисперсионный анализ					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
Регрессия	2	0,234511	0,117255	219,4951	2,92E-15
Остаток	22	0,011752	0,000534		
Итого	24	0,246263			

Таблица 15.

	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>	<i>Нижние 95,0%</i>	<i>Верхние 95,0%</i>
Y-пересечение	-0,63377	0,073039	-8,67723	1,49E-08	-0,78525	-0,4823	-0,78525	-0,4823
Переменная X 1	-0,00055	0,000157	-3,51286	0,001963	-0,00088	-0,00023	-0,00088	-0,00023
Переменная X 6	0,020747	0,00099	20,95189	5,03E-16	0,018693	0,0228	0,018693	0,0228

Исходя из расчетов, можем записать уравнение регрессии таким образом:

$$y = -0,63377 - 0,00055 * x_1 + 0,020747 x_6$$

Оценим качество уравнений через среднюю ошибку аппроксимации:

Таблица 16.

Страна	Y	X1	X6	Теоретическое значение	$\frac{y_i - \hat{y}_i}{y_i}$
Австрия	0,904	115	77	0,900499	0,003873
Австралия	0,922	123	78,2	0,9209954	0,00109
Белоруссия	0,763	74	68	0,736326	0,034959
Бельгия	0,923	111	77,2	0,9068484	0,017499
Великобритания	0,918	113	77,2	0,9057484	0,013346
Германия	0,906	110	77,2	0,9073984	0,001543
Дания	0,905	119	75,7	0,8713279	0,037207
Индия	0,545	146	62,6	0,5846922	0,07283
Испания	0,894	113	78	0,922346	0,031707
Италия	0,9	108	78,2	0,9292454	0,032495
Канада	0,932	113	79	0,943093	0,011902
Казахстан	0,74	71	67,6	0,7296772	0,01395
Китай	0,701	210	69,8	0,6988706	0,003038
Латвия	0,744	94	68,4	0,7336248	0,013945
Нидерланды	0,921	118	77,9	0,9175213	0,003777
Норвегия	0,927	130	78,1	0,9150707	0,012869
Польша	0,802	127	72,5	0,8005375	0,001824
Россия	0,747	61	66,6	0,7144302	0,043601
США	0,927	117	76,7	0,8931749	0,036489
Украина	0,721	46	68,8	0,7683236	0,065636
Финляндия	0,913	107	76,8	0,9007496	0,013418
Франция	0,918	110	78,1	0,9260707	0,008792
Чехия	0,833	99	73,9	0,8449833	0,014386
Швейцария	0,914	101	78,6	0,9413942	0,029972
Швеция	0,923	105	78,5	0,9371195	0,015297
A,%					2,141771