МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА

Институт информационных технологий и технологического образования

Кафедра компьютерные технологии и электронного обучения

Основная профессиональная образовательная программа

Направление подготовки 09.03.01 Информатика и вычислительная техника

Направленность (профиль) «Технологии разработки программного обеспечения»

форма обучения – очная

ЛАБОРАТОРНАЯ РАБОТА №7.1

по дисциплине: «Анализ данных и основы Data science»

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ. ВЫЧИСЛЕНИЕ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

Руководитель:

кандидат педагогических наук, доцент,

Светлана Викторовна Гончарова

Автор работы студент 2 курса

1 группы 1 подгруппы

Чирцов Тимофей Александрович

Санкт-Петербург
2023

Цель работы: провести вычисления коэффициентов корреляции, ранговой корреляции и линейной корреляции.

Оборудование: ПК, Excel

Математические модели:

$$d_i = x_i - y_i$$

$$t = |r_s| \sqrt{\frac{n-2}{1-r_s^2}}$$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Задание 2 (пример 1)

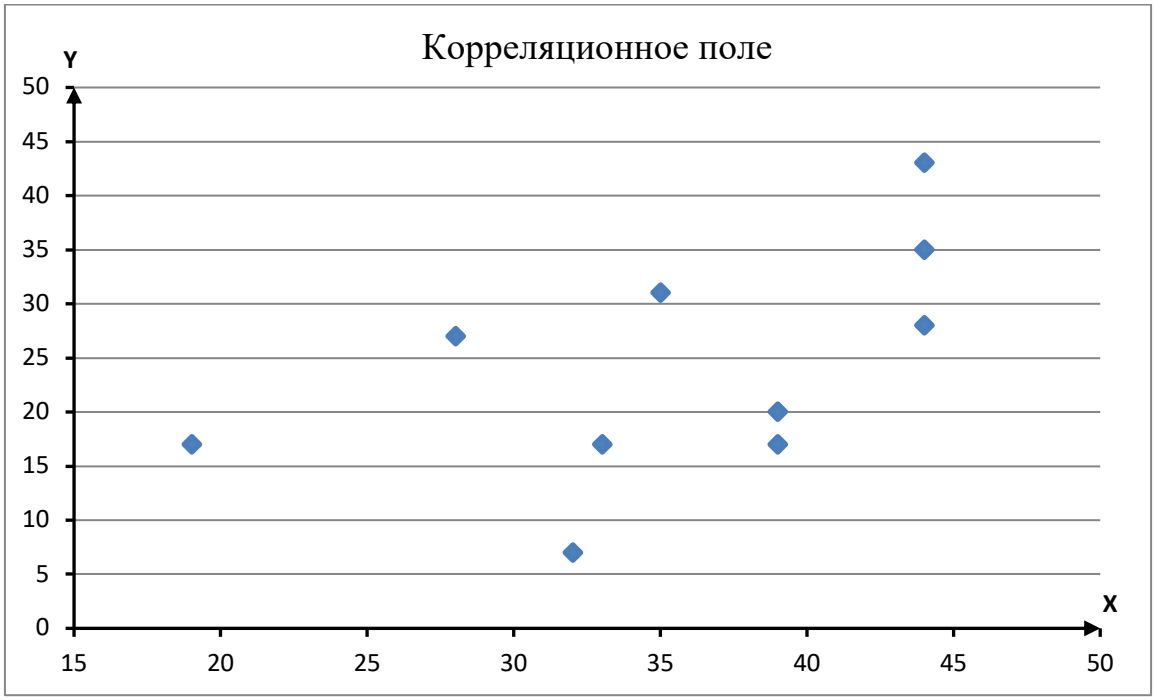
Построить корреляционное поле. Сделать предположение о форме и направлении взаимосвязи двух исследуемых показателей. Вычислить коэффициент линейной корреляции Пирсона и коэффициент ранговой корреляции Спирмена (Примеры 1 и 2 из материалов лекции).

Пример 1.
10 школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Исследователя интересует вопрос: существует ли взаимосвязь между временем решения этих задач?
Переменная X — обозначает среднее время решения наглядно-образных, а переменная Y— среднее время решения вербальных заданий тестов.

№ испытуемых	X	Y
1	19	17
2	32	7
3	33	17
4	44	28
5	28	27
6	35	31
7	39	20
8	39	17
9	44	35
10	44	43

Результат работы:

	A	B	C	D	E	F	G	H	I
1	Пример 1								
2	№ испытуемых	Xi	Yi		X	Y		X	Y
3	1	19	17		9,5	8,5		35,7	24,2
4	2	32	7		16	3,5			
5	3	33	17		16,5	8,5			
6	4	44	28		22	14			
7	5	28	27		14	13,5			
8	6	35	31		17,5	15,5			
9	7	39	20		19,5	10			
10	8	39	17		19,5	8,5			
11	9	44	35		22	17,5			
12	10	44	43		22	21,5			
13									
14	№	(xi-x)*(yi-y)	(xi-x)^2	(yi-y)^2		№	d^2		
15	1	120,24	278,89	51,84		1	4		
16	2	63,64	13,69	295,84		2	625		
17	3	19,44	7,29	51,84		3	256		
18	4	31,54	68,89	14,44		4	256		
19	5	-21,56	59,29	7,84		5	1		
20	6	-4,76	0,49	46,24		6	16		
21	7	-13,86	10,89	17,64		7	361		
22	8	-23,76	10,89	51,84		8	484		
23	9	89,64	68,89	116,64		9	81		
24	10	156,04	68,89	353,44		10	1		
25	сумма	416,6	588,1	1007,6		сумма	2085		
26									
27									
28	r_{xy}	0,54118979		k	8				
29	rs	-11,636364		$r_{крит}$	0,63				
30									



Вывод: предположительно это линейная, положительно направленная связь. Т.к. $R_{xy} < R_{крит}$, следовательно, гипотеза H_1 отвергается и принимается гипотеза H0, иными словами, связь между временем решения

Задание 2 (пример 2)

Пример 2.

Преподавателю и студенту было предложено расположить 10 профессий в порядке их общественной значимости. Ответы перечислены в таблице 1:

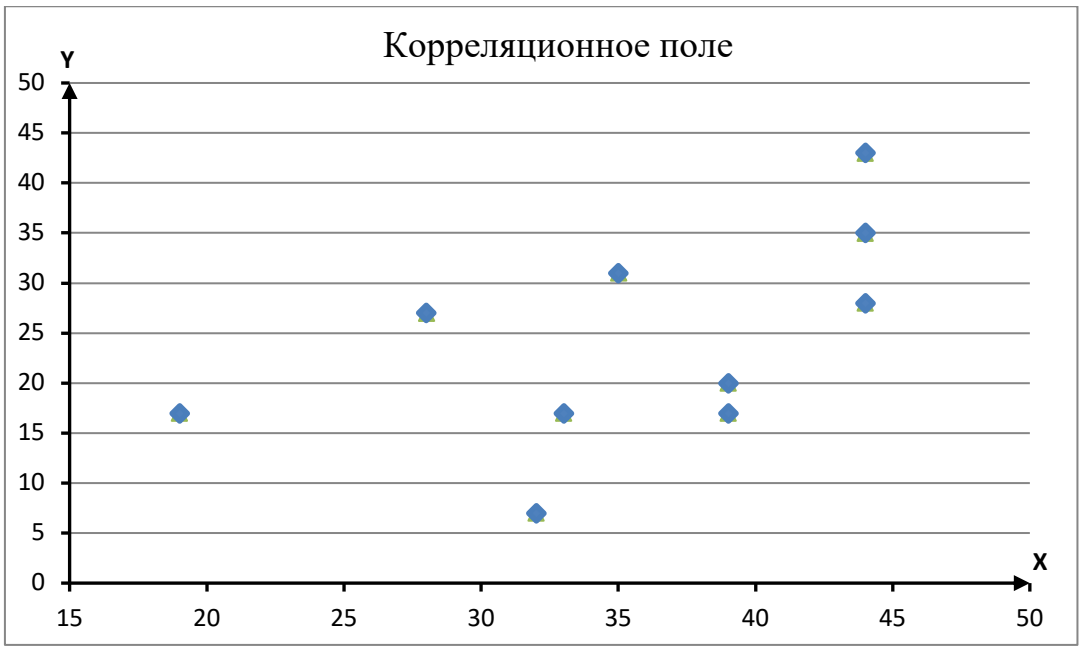
Таблица 1

Оценка преподавателя, x_i	**Профессии**	**Оценка студента, y_i**
3	профессор	2
1	врач	1
4	учитель школы	7
2	Директор магазина	4
8	бухгалтер	5
6	банкир	3
9	водитель	9
5	журналист	8
10	ди-джей	10
7	программист	6

Какова корреляция рангов между двумя рядами оценок ? Одинаково ли мнение преподавателя и студента по этому вопросу ?

Результат работы:

32	Пример 2						
33							
34	Оценка преподавателя, xi	Профессии	Оценка студента, yi		№	di	di^2
35	3	профессор	2		1	1	1
36	1	врач	1		2	0	0
37	4	учитель шко	7		3	-3	9
38	2	директор ма	4		4	-2	4
39	8	бухгалтер	5		5	3	9
40	6	банкир	3		6	3	9
41	9	водитель	9		7	0	0
42	5	журналист	8		8	-3	9
43	10	ди-джей	10		9	0	0
44	7	программист	6		10	1	1
45					сумма	0	42
46							
47	rs	0,7455					
48							
49	t	3,16365285					



Вывод: Гипотеза: существует линейная связь с положительным направлением. Значит, на уровне значимости 5% линейная связь между мнениями студента и преподавателя является статистически значимой, а также она является положительно направленной и имеет сильную тесноту связи. $t_{расч} > t_{кр}$ (4,653 > 1,86),

Задание 3.1

Задание 3. Построить корреляционное поле. Сделать предположение о форме и направлении взаимосвязи двух исследуемых показателей. Найти значения коэффициентов ранговой корреляции Спирмена.

Задача 3.1 С помощью коэффициента ранговой корреляции установить зависимость между стажем практической работы и временем решения контрольной задачи у 10 программистов на основе следующих данных:

Номера испытуемых	1	2	3	4	5	6	7	8	9	10
Стаж (в мес).	32	15	16	18	20	28	21	29	23	17
Время решения (в мин.)	12	24	23	21	20	9	11	10	15	16

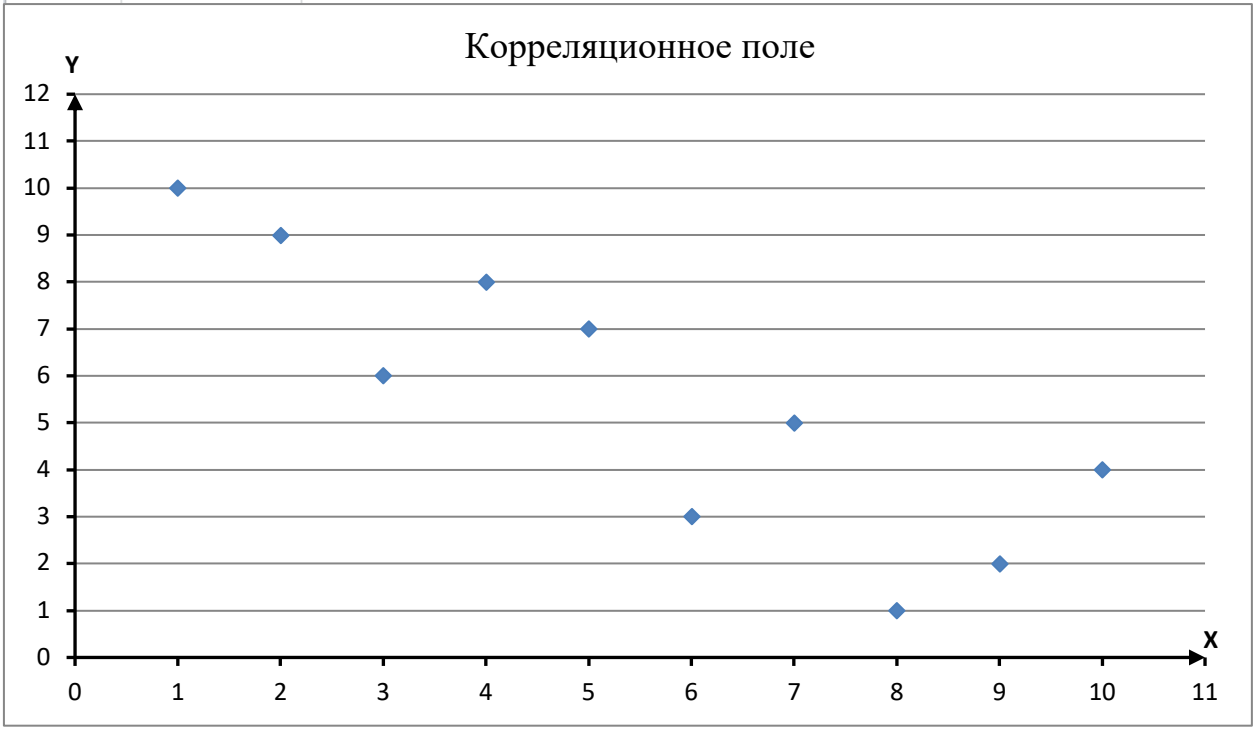
Результат работы:

Номера испытуемых	стаж	Время решения	Ранжирование стажа(Xi)	Ранжировка времени (Yi)
1	32	12	10	4
2	15	24	1	10
3	16	23	2	9
4	18	21	4	8
5	20	20	5	7
6	28	9	8	1
7	21	11	6	3
8	29	10	9	2
9	23	15	7	5
10	17	16	3	6

		di	di^2
15			
16		6	36
17		-9	81
18		-7	49
19		-4	16
20		-2	4
21		7	49
22		3	9
23		7	49
24		2	4
25		-3	9
26	Сумма	0	306

rs	-0,854545455

Ткр = T0,05, 8	1,86
t	4,653693452



Вывод: Гипотеза: существует линейная связь с отрицательным направлением и сильной теснотой связи. Значит на уровне значимости 5% линейная связь между стажем и временем решения задачи является статистически значимой, а также она является отрицательно направленной и имеет сильную тесноту связи (-1 < -0,855 <-0,7).

Задание 3.2

Задача 3.2.

Три арбитра оценили мастерство 10 спортсменов, в итоге были получены три последовательности рангов (в первой строке приведены ранги арбитра A, во второй – ранги арбитра B, в третьей – ранги арбитра C):

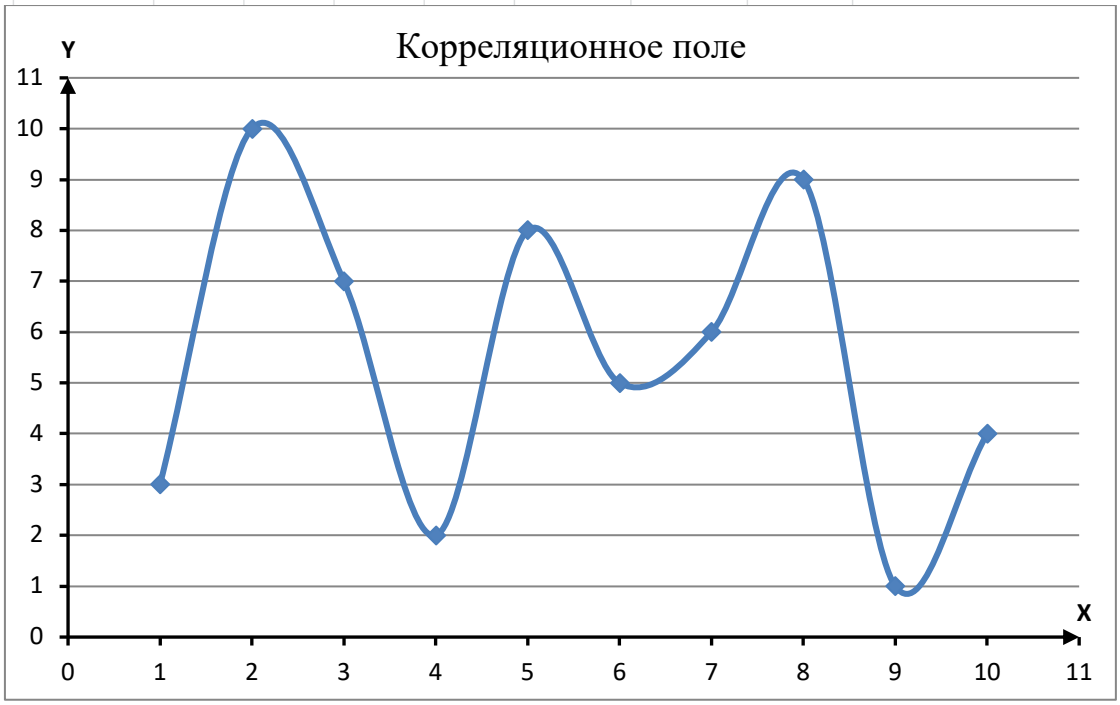
x_i	1	2	3	4	5	6	7	8	9	10
y_i	3	10	7	2	8	5	6	9	1	4
z_i	6	2	1	3	9	4	5	7	10	8

Определить пару арбитров, оценки которых наиболее согласуются, используя коэффициент ранговой корреляции Спирмена. Построить корреляционное поле.

Результат работы:

Пример 3.2							
№	xi	yi	zi		di^2 = Xi - Yi	di ^2= Xi - Zi	di^2 = Yi-Zl
1	1	3	6		4	25	9
2	2	10	2		64	0	64
3	3	7	1		16	4	36
4	4	2	3		4	1	1
5	5	8	9		9	16	1
6	6	5	4		1	4	1
7	7	6	5		1	4	1
8	8	9	7		1	1	4
9	9	1	10		64	1	81
10	10	4	8		36	4	16
				СУММ	200	60	214

rs (Xi Yi)	-0,2121
rs (Xi Zi)	0,63636
rs (Yi Zi)	-0,297



Вывод: Оценки больше всего согласуются у арбитров A и C.

Задание 4.1

Задание 4.
Задача 4.1. Построить корреляционное поле. Сделать предположение о форме и направлении взаимосвязи двух исследуемых показателей. Найти значение коэффициентов линейной корреляции.
Необходимо определить взаимосвязь характеристик: агрессивности и IQ у школьников по полученным данным тестирования.

№ п/п	Данные по	Данные по IQ

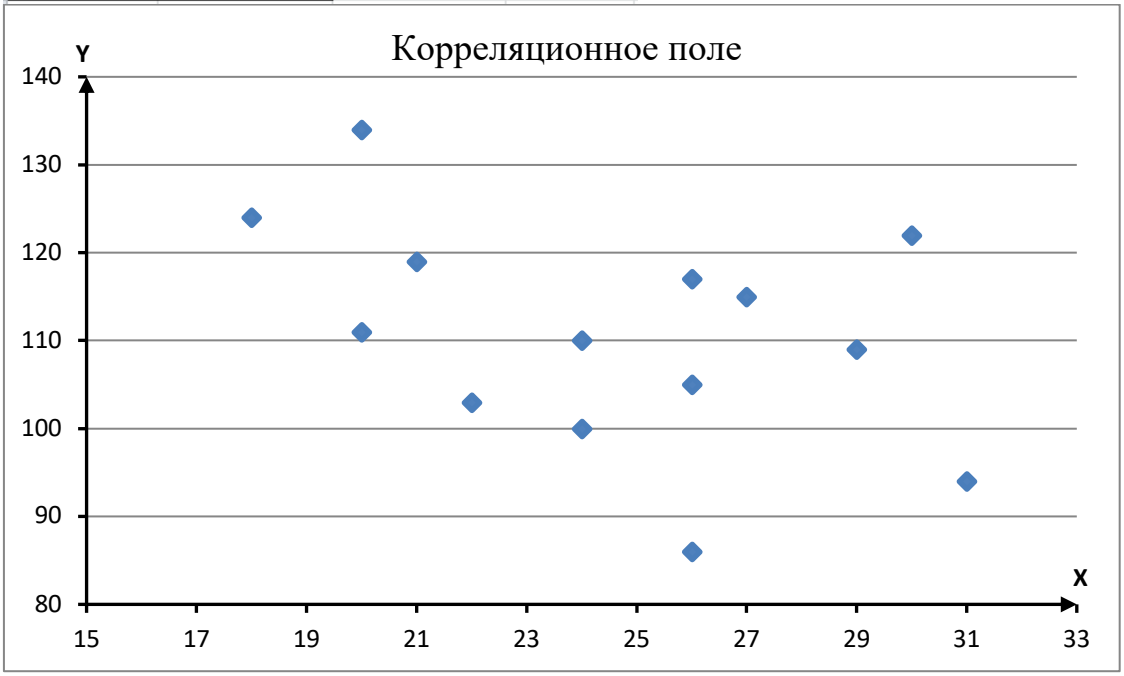
	агрессивности X_{agr}	Y_{IQ}
1	24	100
2	27	115
3	26	117
4	21	119
5	20	134
6	31	94
7	26	105
8	22	103
9	20	111
10	18	124
11	30	122
12	29	109
13	24	110
14	26	86

Результат работы:

№ п/п	данные по агрессивности X agr	Данные по IQ Y
1	24	100
2	27	115
3	26	117
4	21	119
5	20	134
6	31	94
7	26	105
8	22	103
9	20	111
10	18	124
11	30	122
12	29	109
13	24	110
14	26	86
ср зна	24,57142857	110,6428571

№	(xi-x)*(yi-y)	(xi-x)^2	(yi-y)^2
1	6,081632653	0,326530612	113,2704
2	10,58163265	5,897959184	18,98469
3	9,081632653	2,040816327	40,41327
4	-29,8469388	12,75510204	69,84184
5	-106,77551	20,89795918	545,5561
6	-106,989796	41,32653061	276,9847
7	-8,06122449	2,040816327	31,84184
8	19,65306122	6,612244898	58,41327
9	-1,63265306	20,89795918	0,127551
10	-87,7755102	43,18367347	178,4133
11	61,65306122	29,46938776	128,9847
12	-7,2755102	19,6122449	2,69898
13	0,367346939	0,326530612	0,413265
14	-35,2040816	2,040816327	607,2704
Сумм	-276,142857	207,4285714	2073,214

r xy	-0,42109243
r крит	0,53



Вывод: $R_{крит} > |R_{xy}|$, а значит, гипотеза H_0 принимается.

Задание 4.2

Задача 4.2

На основании наблюдений за развивающимся сайтом и изменением его средневзвешенной позиции по основным запросам в поисковой системе необходимо проверить, можно ли говорить о линейной зависимости между позицией сайта и числом посетителей. Построить корреляционное поле.

Исходные данные:
X - число посетителей в сутки;
Y – усредненная позиция сайта в поисковой системе.
В таблице даны значения признаков.

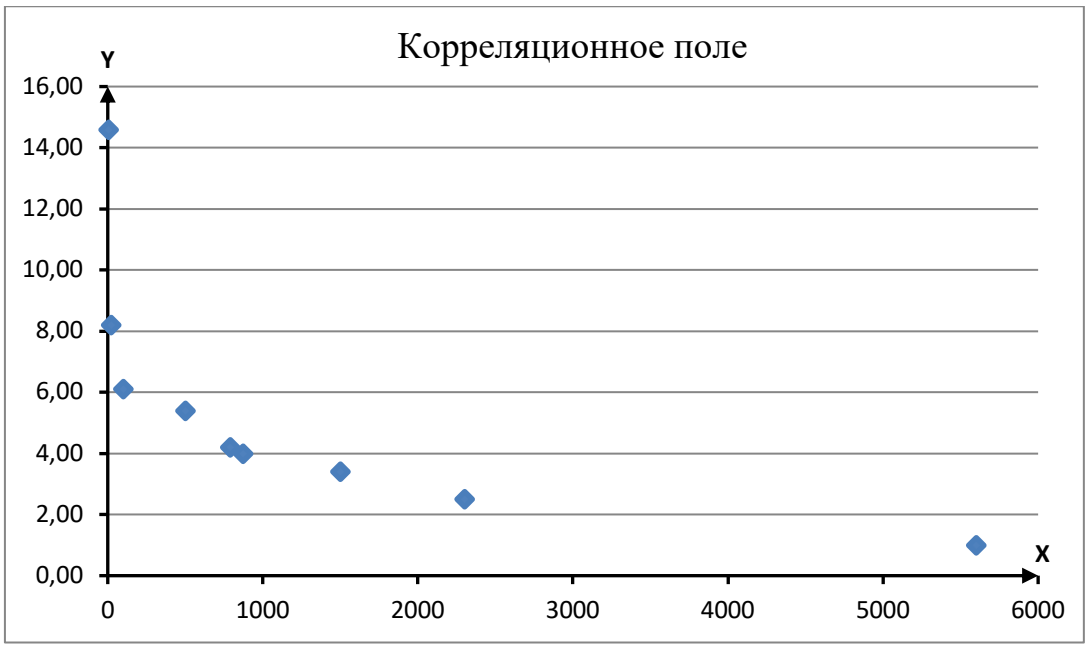
№ п/п	Число посетителей в сутки, X	Усредненная позиция сайта в поисковой системе, Y
1	500	5.4
2	790	4.2
3	870	4.0
4	1500	3.4
5	2300	2.5
6	5600	1.0
7	100	6.1
8	20	8.2
9	5	14.6

Результат работы:

№	число посетителей в сутки , X	Усредненная позиция сайта в поисковой системе, Y
1	500	5,40
2	790	4,20
3	870	4,00
4	1500	3,40
5	2300	2,50
6	5600	1,00
7	100	6,10
8	20	8,20
9	5	14,60
Ср знач	1298,333333	5,49

№	(xi-x)*(yi-y)	(xi-x)^2	(yi-y)^2
1	71,85	637336,1111	0,01
2	655,75	258402,7778	1,66
3	638,2166667	183469,4444	2,22
4	-421,483333	40669,44444	4,37
5	-2994,98333	1003336,111	8,94
6	-19314,4833	18504336,11	20,16
7	-730,983333	1436002,778	0,37
8	-3464,28333	1634136,111	7,34
9	-11782,2667	1672711,111	82,99
сумм	-37342,6667	25370400	128,0689

r xy	-0,65511845
r крит	0,67



Вывод: $R_{крит} > |R_{xy}|$, а значит гипотеза H_0 принимается. Возможно существует линейная связь с отрицательным направлением и средней теснотой связи.

Вывод по лабораторной работе: с помощью электронных таблиц Excel нам удалось провести вычисления коэффициентов корреляции, ранговой корреляции и линейной корреляции.