МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А.И. ГЕРЦЕНА»

Направление подготовки

09.03.01 – Информатика и вычислительная техника

Профиль «Технологии разработки программного обеспечения»

Индивидуальная задача по дисциплине Анализ данных и основы Data Science

Работу выполнил студент 2 курса 2-1 группы:

Воложанин Владислав Олегович

САНКТ-ПЕТЕРБУРГ

2023

Задача

Был проведён опрос среди группы из 20 человек из разных сфер жизни. Выборка состоит из:

1. Индекс счастья (по шкале от 1 до 10) - это оценка уровня счастья и удовлетворенности жизни, где 1 - очень низкий уровень счастья, а 10 - очень высокий уровень счастья.

2. Уровень дохода.

3. Качество образования (по шкале от 1 до 5) - это индивидуальная оценка качества образования человека, где 1 - низкое качество образования, а 5 - высокое качество образования.

4. Доступ к медицинскому обслуживанию (по шкале от 1 до 5) - это оценка доступности и качества медицинских услуг, которые получает человек, где 1 - низкий уровень доступа к медицине, а 5 - высокий уровень доступа к медицине.

5. Стаж работы.

Задачи для исследования

1. Расчёт коэффициенты корреляции между индекс счастья и уровнем дохода.
2. Расчёт коэффициенты корреляции между индекс счастья и качеством образования.
3. Построить корреляционные поля.
4. Сделать статическую проверку значимости коэффициентов корреляций.
5. Реализовать регрессионный анализ.

Человек	Индекс счастья	Уровень дохода	Качество образования	Доступ к мед. обслуживанию	Стаж работы в нед.
1	7	50000	4	3	52
2	6	40000	3	2	39
3	8	60000	5	4	68
4	5	30000	2	1	26
5	9	80000	5	5	78
6	4	20000	1	1	13
7	7	55000	4	3	56
8	6	45000	3	2	43
9	8	65000	5	4	61
10	5	35000	2	1	30
11	9	75000	5	5	71
12	4	25000	1	1	18
13	7	52000	4	3	54
14	6	42000	3	2	41
15	8	62000	5	4	64
16	5	32000	2	1	28
17	9	72000	5	5	74
18	4	22000	1	1	16
19	7	57000	4	3	58
20	6	47000	3	2	45

Корреляционный анализ

Коэффициент линейной корреляции Пирсона:

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

Интерпретация коэффициентов корреляции.

1	$\lvert r\rvert = 1$	функциональная зависимость
2	$0{,}7 \le \lvert r\rvert \le 0{,}99$	сильная статистическая взаимосвязь
3	$0{,}5 \le \lvert r\rvert \le 0{,}69$	средняя статистическая взаимосвязь
4	$0{,}2 \le \lvert r\rvert \le 0{,}49$	слабая статистическая взаимосвязь
5	$0{,}09 \le \lvert r\rvert \le 0{,}19$	очень слабая статистическая взаимосвязь
6	$\lvert r\rvert = 0$	**корреляции нет (линейной)**

Расчёт коэффициента корреляции между индексом счастья и уровнем дохода

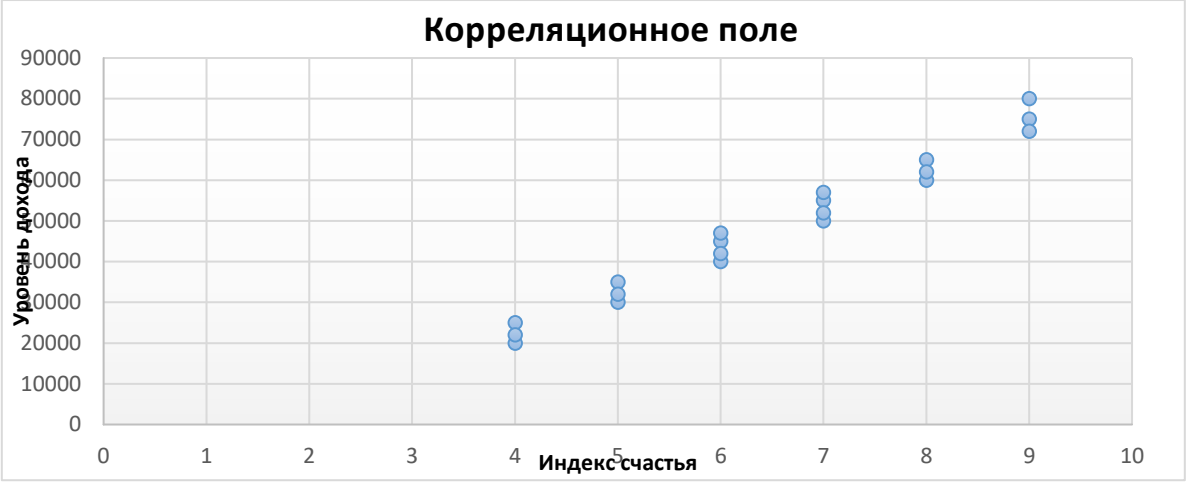
	B	C	D	E	F
25	Рассчитаем коэффициент корреляции между индексом счастья и уровнем дохода.				
26	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
27	1	0	1700	2890000	850
28	-1	0	-8300	68890000	4150
29	2	2	11700	136890000	17550
30	-2	2	-18300	334890000	27450
31	3	6	31700	1004890000	79250
32	-3	6	-28300	800890000	70750
33	1	0	6700	44890000	3350
34	-1	0	-3300	10890000	1650
35	2	2	16700	278890000	25050
36	-2	2	-13300	176890000	19950
37	3	6	26700	712890000	66750
38	-3	6	-23300	542890000	58250
39	1	0	3700	13690000	1850
40	-1	0	-6300	39690000	3150
41	2	2	13700	187690000	20550
42	-2	2	-16300	265690000	24450
43	3	6	23700	561690000	59250
44	-3	6	-26300	691690000	65750
45	1	0	8700	75690000	4350
46	-1	0	-1300	1690000	650

=СУММ(F27:F46)/КОРЕНЬ((СУММ(C27:C46)*СУММ(E27:E46)))

$r_s =$	0,987969049

Коэффициент корреляции между индексом счастья и уровнем дохода равен 0,987969049.

Это говорит о том, что существует сильная положительная корреляция между этими двумя переменными. Также это можно также наблюдать и на корреляционном поле.



Статическая проверка значимости коэффициента корреляции

уровень значимости (α) = 0.05.

$$t_{\text{расч}} = |r_s| \cdot \sqrt{\frac{n-2}{1-r_s^2}}$$

Функция **TTEST** – используется для выполнения двустороннего теста на равенство средних значений двух выборок. Она помогает определить, существует ли статистически значимая разница между средними значениями двух групп данных.

=TTECT(C3:C22;D3:D22;2;1)

p	1,95501E-10

= 0.000000000195501.

Значение *p* гораздо меньше стандартного уровня значимости 0.05. Следовательно, можно сделать вывод, что связь между индексом счастья и уровнем дохода является статистически значимой.

Расчёт коэффициента корреляции между между индексом счастья и качеством образования

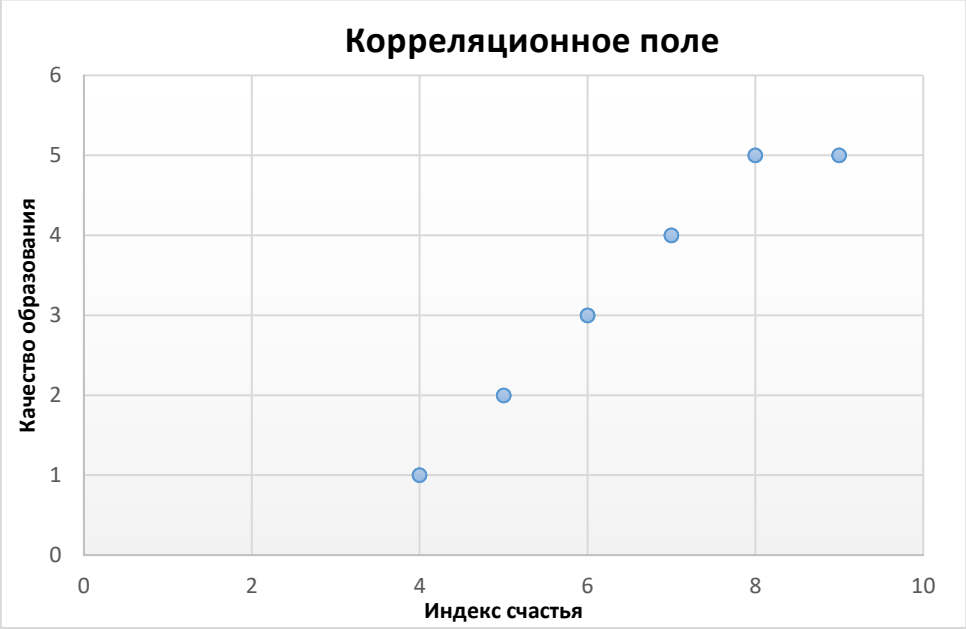
	H	I	J	K	L
25	Рассчитаем коэффициент корреляции между между индексом счастья и качеством образования				
26	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
27	0,5	0,25	0,65	0,4225	0,325
28	-0,5	0,25	-0,35	0,1225	0,175
29	1,5	2,25	1,65	2,7225	2,475
30	-1,5	2,25	-1,35	1,8225	2,025
31	2,5	6,25	1,65	2,7225	4,125
32	-2,5	6,25	-2,35	5,5225	5,875
33	0,5	0,25	0,65	0,4225	0,325
34	-0,5	0,25	-0,35	0,1225	0,175
35	1,5	2,25	1,65	2,7225	2,475
36	-1,5	2,25	-1,35	1,8225	2,025
37	2,5	6,25	1,65	2,7225	4,125
38	-2,5	6,25	-2,35	5,5225	5,875
39	0,5	0,25	0,65	0,4225	0,325
40	-0,5	0,25	-0,35	0,1225	0,175
41	1,5	2,25	1,65	2,7225	2,475
42	-1,5	2,25	-1,35	1,8225	2,025
43	2,5	6,25	1,65	2,7225	4,125
44	-2,5	6,25	-2,35	5,5225	5,875
45	0,5	0,25	0,65	0,4225	0,325
46	-0,5	0,25	-0,35	0,1225	0,175

=СУММ(L27:L46)/КОРЕНЬ((СУММ(I27:I46)*СУММ(K27:K46)))

$r_s =$	0,981472267

Коэффициент корреляции между индексом счастья и качеством образования составляет 0,981472267.

Это указывает на сильную положительную связь между этими двумя переменными. Также это можно также наблюдать и на корреляционном поле.



Статическая проверка значимости коэффициента корреляции

уровень значимости (α) = 0.05.

=ТТЕСТ(C3:C22;E3:E22;2;1)

p	1,74232E-19

$= 0.00000000000000000017$.

Значение *p* гораздо меньше стандартного уровня значимости 0.05. Следовательно, можно сделать вывод, что связь между индексом счастья и качеством образования является статистически значимой.

Регрессионный анализ

Коэффициенты линейной регрессии вычисляются по формулам:

$$a = \frac{n \cdot \sum xy - \sum x \cdot \sum y)}{n \cdot \sum x^2 - (\sum x)^2}$$

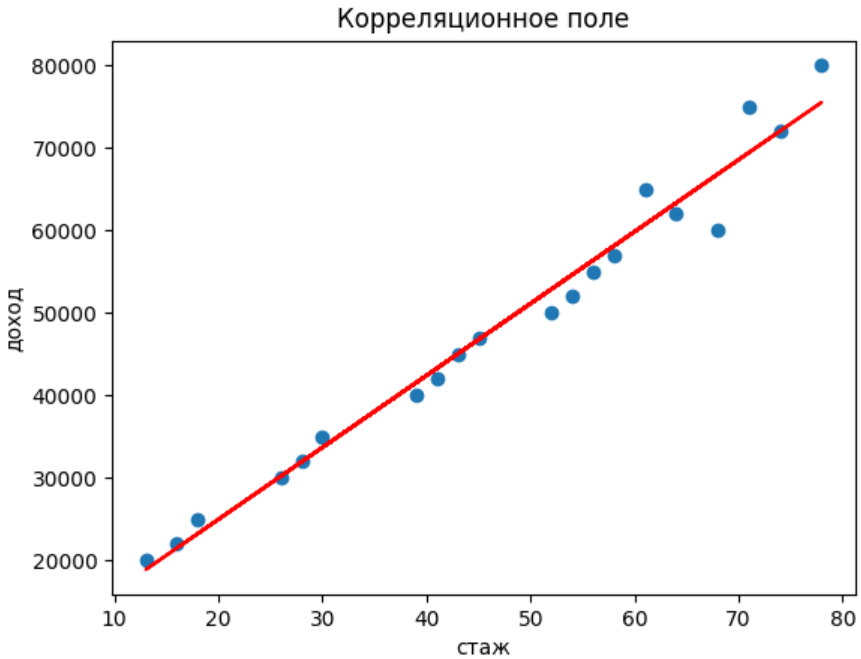
$$b = \frac{\sum y - a \cdot \sum x}{n}$$

```python
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import t, f
def regression_model(x, y):
    # Проверяем, что массивы x и y имеют одинаковую длину
    if len(x) != len(y):
        raise ValueError("Массивы x и y должны иметь одинаковую длину")
    # Количество точек
    n = len(x)
    # Вычисляем суммы значений x, y и их произведений
    sum_x = np.sum(x)
    sum_y = np.sum(y)
    sum_xy = np.sum(x * y)
    sum_x_squared = np.sum(x ** 2)
    # Вычисляем коэффициенты a и b
    a = (n * sum_xy - sum_x * sum_y) / (n * sum_x_squared - sum_x ** 2)
    b = (sum_y - a * sum_x) / n
    # Строим график
    plt.scatter(x, y)
    plt.plot(x, a * x + b, 'r')
    plt.xlabel('стаж')
    plt.ylabel('доход')
    plt.title('Регрессивная модель')
    plt.show()
    return a, b
def main():
    y = np.array([50000, 40000, 60000, 30000, 80000, 20000, 55000,
    45000, 65000, 35000, 75000, 25000, 52000, 42000, 62000, 32000, 72000,
    22000, 57000, 47000])
    x = np.array([52, 39, 68, 26, 78, 13, 56,
    43, 61, 30, 71, 18, 54, 41, 64, 28, 74, 16, 58,
    45])
    predict = regression_model(x, y)
if __name__ == '__main__':
    main()
```

a = 872.1474642307567

b = 7527.106047212123

Результат программы:



Итоговый вывод

Оба коэффициента корреляции указывают на сильную связь между индексом счастья и уровнем дохода, а также между индексом счастья и качеством образования. Повышение дохода и получение лучшего образования могут положительно влиять на уровень счастья людей. Также связь между индексом счастья и уровнем дохода, и связь между индексом счастья и качеством образования является статистически значимыми.

Программа реализует метод наименьших квадратов для построения линейной регрессионной модели. Она принимает на вход два массива данных: x (стаж) и y (доход), соответствующие значениям независимой и зависимой переменных соответственно. Вначале программа проверяет, что длины массивов x и y совпадают. Затем вычисляет коэффициенты a и b линейной регрессионной модели.