Выборочное среднее для X

В этом шаге вычисляется выборочное среднее для переменной X. Формула для выборочного среднего выглядит следующим образом:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

где:

- $\overline{X}$ - выборочное среднее для X.
- n - количество наблюдений.
- x_i - значения переменной X.

```
x_mean = data['X'].mean()
print(f'Выборочное среднее для X: {x_mean:.2f}')
```

Выборочное среднее для Y

Аналогично, в этом шаге вычисляется выборочное среднее для переменной Y, используя формулу:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

где:

- $\overline{Y}$ - выборочное среднее для Y.
- n - количество наблюдений.
- y_i - значения переменной Y.

```
y_mean = data['Y'].mean()
print(f'Выборочное среднее для Y: {y_mean:.2f}')
```

Выборочное среднее для X и Y

Этот шаг вычисляет выборочное среднее для произведения переменных X и Y. Формула для выборочного среднего произведения двух переменных:

$$\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} (x_i y_i)$$

где:

- $\overline{XY}$ - выборочное среднее для произведения X и Y.
- n - количество наблюдений.
- x_i - значения переменной X.
- y_i - значения переменной Y.

```
xy_mean = (data['X'] * data['Y']).mean()
print(f'Выборочное среднее для XY: {xy_mean:.2f}')
```

Расчет дисперсий и стандартных отклонений

В этом разделе производится расчет дисперсий и стандартных отклонений для переменных X и Y.

Выборочная дисперсия для X

Выборочная дисперсия для переменной X вычисляется с использованием следующей формулы:

$$D_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

где:

- D_x^2 - выборочная дисперсия для X.
- n - количество наблюдений.
- x_i - значения переменной X.
- $\overline{X}$ - выборочное среднее для X.

```
x_var = data['X'].var(ddof=0)
print(f'Выборочная дисперсия для X: {x_var:.2f}')
```

Стандартное отклонение для X

Стандартное отклонение для переменной X рассчитывается как квадратный корень из выборочной дисперсии:

$$\sigma_x = \sqrt{D_x^2}$$

где:

- σ_x - стандартное отклонение для X.
- D_x^2 - выборочная дисперсия для X.

```
x_std = data['X'].std(ddof=0)
```

```
print(f'Стандартное отклонение для X: {x_std:.2f}')
```

Выборочная дисперсия для Y

Аналогично, выборочная дисперсия для переменной Y вычисляется с использованием формулы:

$$D_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2$$

где:

- D_y^2 - выборочная дисперсия для Y.
- n - количество наблюдений.
- y_i - значения переменной Y.
- $\overline{Y}$ - выборочное среднее для Y.

```
y_var = data['Y'].var(ddof=0)
print(f'Выборочная дисперсия для Y: {y_var:.2f}')
```

Стандартное отклонение для Y

Стандартное отклонение для переменной Y рассчитывается как квадратный корень из выборочной дисперсии:

$$S_y = \sqrt{D_y^2}$$

где:

- S_y - стандартное отклонение для Y.
- D_y^2 - выборочная дисперсия для Y.

```
s_std = data['Y'].std(ddof=0)
print(f'Стандартное отклонение для Y: {s_std:.2f}')
```

Расчет коэффициента корреляции

Этот шаг включает в себя вычисление коэффициента корреляции между переменными X и Y. Коэффициент корреляции измеряет степень линейной зависимости между переменными и может принимать значения от -1 до 1.

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

где n - количество наблюдений, x_i и y_i - значения случайных величин, $\bar{x}$ и $\bar{y}$ - средние значения.

$$r_{XY} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sigma_x \cdot \sigma_y}$$

```
r_xy = (xy_mean - x_mean * y_mean) / (x_std * s_std)
print(f'Коэффициент корреляции: {r_xy}')
```

Расчет параметров a и b

В этом разделе вычисляются параметры a и b для линейного уравнения регрессии.

Параметр a

Параметр a (наклон) рассчитывается с использованием следующей формулы:

$$a = \frac{S_y}{S_x} \cdot r_{XY}$$

где:

- a - параметр наклона.
- S_y - стандартное отклонение для Y.
- S_x - стандартное отклонение для X.
- r_{XY} - коэффициент корреляции между X и Y.

```
a = s_std / x_std * r_xy
print(f'Параметр a: {a:.2f}')
```

Параметр b

Параметр b (пересечение) рассчитывается с использованием следующей формулы:

$$b = \overline{Y} - a \cdot \overline{X}$$

где:

- b - параметр пересечения.
- $\overline{Y}$ - выборочное среднее для Y.
- $\overline{X}$ - выборочное среднее для X.
- a - параметр наклона.

```
b = y_mean - a * x_mean
print(f'Параметр b: {b:.2f}')
```

Проверка гипотезы о значимости коэффициента корреляции

В этом шаге проводится статистический тест для проверки значимости коэффициента корреляции r_{XY}.

Гипотезы

- **Нулевая гипотеза (H0):** $r_{XY} = 0$ - это нулевая гипотеза, которую мы хотим проверить. Она утверждает, что коэффициент корреляции между X и Y равен нулю (нет линейной зависимости).
- **Альтернативная гипотеза (H1):** $r_{XY} \neq 0$ - это альтернативная гипотеза, которая утверждает, что коэффициент корреляции между X и Y не равен нулю (существует линейная зависимость).

Расчетный критерий Стьюдента

Для проверки гипотезы используется t-критерий Стьюдента для коэффициента корреляции r_{XY}:

$$t_{XY} = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

где:

- t_{XY} - расчетное значение t-критерия Стьюдента.
- r_{XY} - коэффициент корреляции между X и Y.
- n - количество наблюдений.

Python-код для вычисления расчетного значения t-критерия и вывода результата:

```
t_value = abs(r_xy) * np.sqrt(len(data['X']) - 2) / np.sqrt(1 - r_xy ** 2)
print(f'Значение критерия: {t_value:.2f}')
```

Критическое значение критерия Стьюдента

Для заданного уровня значимости α и количества степеней свободы $(n-2)$ вычисляется критическое значение t-критерия Стьюдента:

$$t_\alpha = t_{1-\frac{\alpha}{2}, n-2}$$

где:

- t_α - критическое значение t-критерия Стьюдента.
- α - уровень значимости.

- n - количество наблюдений.

Python-код для вычисления критического значения критерия и вывода результата:

```
t_alpha = t.ppf(1 - alpha / 2, len(data['X']) - 2)
print(f'Критическое значение критерия: {t_alpha:.2f}')
```

Вывод о значимости коэффициента корреляции

На последнем этапе производится сравнение расчетного значения t-критерия с критическим

значением. Если $|t_{XY}| < t_\alpha$, то гипотеза H_0 (отсутствие линейной зависимости) принимается, иначе гипотеза H_0 отвергается, что означает, что между переменными X и Y существует статистически значимая линейная зависимость.

Python-код для вывода результата:

```
if abs(t_value) < t_alpha:
    display(Markdown('Коэффициент корреляции значимый, гипотеза $H_0$
принимается, т.к. $t_{эмп} < t_{кр}$.'))
else:
    display(Markdown('Коэффициент корреляции не значимый'))
```

Линейные уравнения регрессии

В заключительной части проводится анализ линейной регрессии между переменными X и Y. Для этого вычисляются два линейных уравнения регрессии: одно для Y на X и другое для X на Y.

Линейное уравнение регрессии Y на X

Линейное уравнение регрессии Y на X имеет следующий вид:

$$Y = aX + b$$

где:

- Y - зависимая переменная.
- X - независимая переменная.
- a - параметр наклона (slope).
- b - параметр пересечения (intercept).

Python-код для вычисления параметров a и b и вывода уравнения регрессии:

```
slope_yx, intercept_yx, r_value_yx, p_value_yx, std_err_yx =
linregress(data['Y'], data['X'])
```

```python
print("Линейное уравнение регрессии Y на X: Y = {:.2f}X +
{:.2f}".format(slope_yx, intercept_yx))
```

Линейное уравнение регрессии X на Y

Линейное уравнение регрессии X на Y имеет следующий вид:

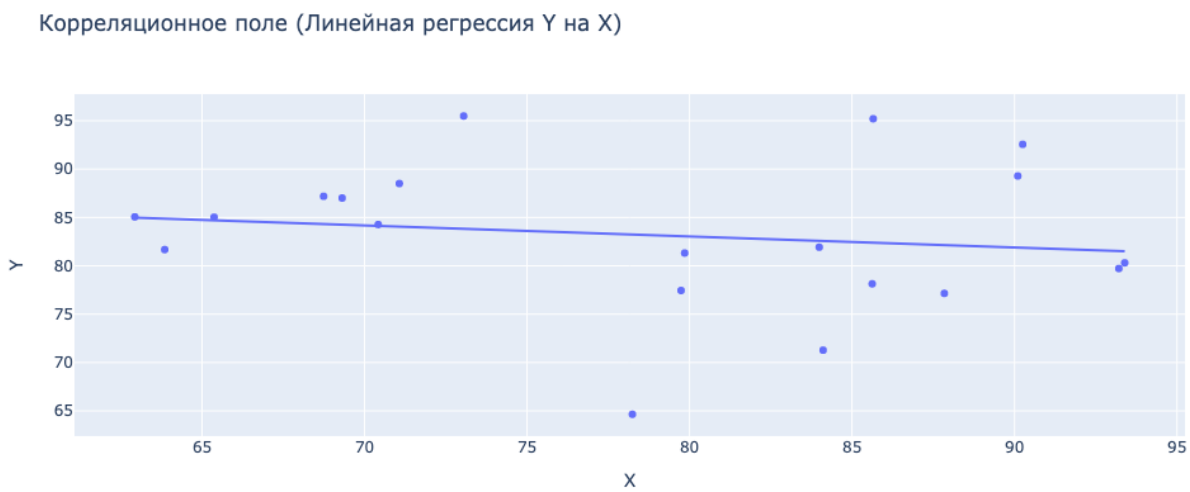
$$X = aY + b$$

где:

- X - зависимая переменная.
- Y - независимая переменная.
- a - параметр наклона (slope).
- b - параметр пересечения (intercept).

Python-код для вычисления параметров a и b и вывода уравнения регрессии:

```python
slope_xy, intercept_xy, r_value_xy, p_value_xy, std_err_xy =
linregress(data['X'], data['Y'])
print("Линейное уравнение регрессии X на Y: X = {:.2f}Y +
{:.2f}".format(slope_xy, intercept_xy))
```



Визуализация результата

Визуализация включает построение точечных графиков рассеяния для линейных регрессий Y на X и X на Y, а также одного графика, на котором отображены оба уравнения регрессии. Графики строятся с использованием библиотеки Plotly.

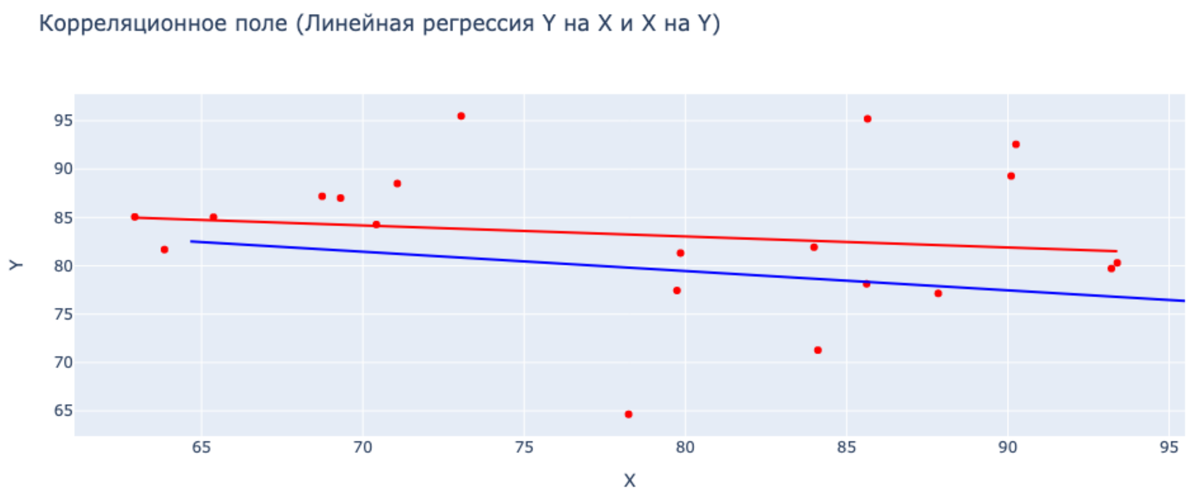
Python-код для построения графиков:

```
fig = px.scatter(data, x='Y', y='X', trendline='ols',
title='Корреляционное поле (Линейная регрессия Y на X)')
fig.show()

px.scatter(data, x='X', y='Y', trendline='ols', title='Корреляционное поле
(Линейная регрессия X на Y)')

fig = px.scatter(data, x='X', y='Y', trendline='ols',
title='Корреляционное поле (Линейная регрессия Y на X и X на Y)',
color_discrete_sequence=['red'])
fig.add_trace(px.scatter(data, x='Y', y='X', trendline='ols',
color_discrete_sequence=['blue']).data[1])

fig.update_layout(showlegend=False)
fig.show()
```



Это завершает описание каждого шага процедуры анализа данных, включая вычисления, проверку гипотезы о значимости коэффициента корреляции и построение линейных уравнений регрессии с визуализацией результатов.
