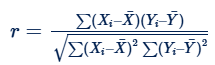
**O que é a Correlação de Pearson?**

A correlação de Pearson, também conhecida como coeficiente de correlação de Pearson, é uma medida estatística que avalia a força e a direção da relação linear entre duas variáveis contínuas. O coeficiente de correlação de Pearson é representado pela letra **r** e varia de -1 a 1.

**Fórmula do Coeficiente de Correlação de Pearson**

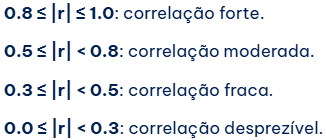
A fórmula para calcular o coeficiente de correlação de Pearson é:

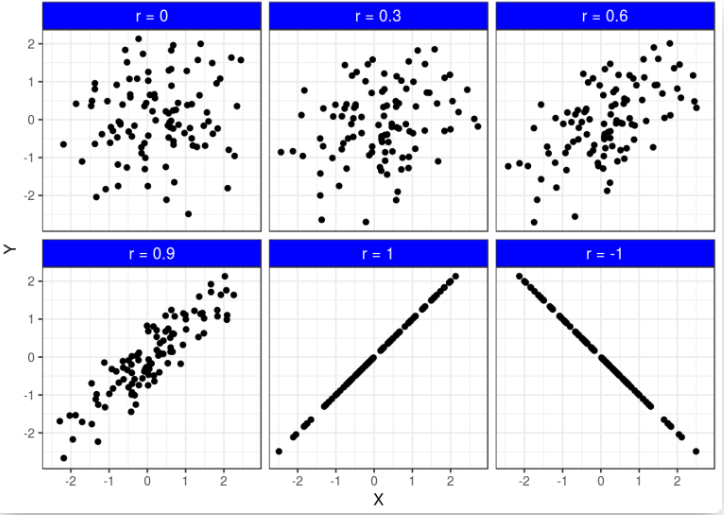


Onde:

* Xi e Yi são os valores individuais das variáveis X e Y.
* X¯ e Y¯ são as médias de X e Y, respectivamente.

**Interpretação dos Valores de r**

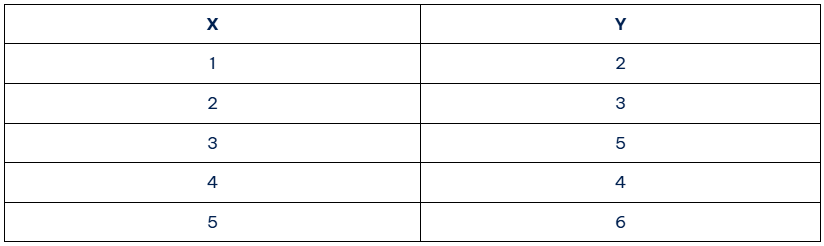
* **r = 1**: Correlação positiva perfeita. À medida que uma variável aumenta, a outra também aumenta.
* **r = -1**: Correlação negativa perfeita. À medida que uma variável aumenta, a outra diminui.
* **r = 0**: Nenhuma correlação linear. As variáveis não têm uma relação linear aparente.
* 



**Passos para Calcular a Correlação de Pearson**

1. **Calcular as Médias**: Calcule a média de cada variável.
2. **Calcular as Diferenças**: Subtraia a média de cada valor individual para obter as diferenças.
3. **Multiplicar as Diferenças**: Multiplique as diferenças correspondentes de X e Y.
4. **Somar os Produtos**: Some todos os produtos das diferenças.
5. **Calcular os Quadrados das Diferenças**: Calcule os quadrados das diferenças para cada variável.
6. **Somar os Quadrados**: Some os quadrados das diferenças para cada variável.
7. **Dividir**: Divida a soma dos produtos das diferenças pela raiz quadrada do produto das somas dos quadrados das diferenças.

**Exemplo Prático**

Vamos supor que temos os seguintes dados para duas variáveis X e Y:  
  


**1.Calcular as Médias**:

* + X = = 3
  + Y = = 4

**2.Em seguida, calculamos os componentes da fórmula**:

* (1−3)(2−4)+(2−3)(3−4)+(3−3)(5−4)+(4−3)(4−4)+(5−3)(6−4)
* 4 + 1 + 0 + 0 + 4
* 9

2.1. **Agora calculamos os denominadores começando pelo X**:

* ∑(Xi–X¯)2
* (1−3)2 +(2−3)2 +(3−3)2 +(4−3)2 +(5−3)2
* 4+1+0+1+4
* 10

2.2. **Agora calculamos o denominador Y:**

* ∑(Yi–Y¯)2
* (2−4)2+(3−4)2+(5−4)2+(4−4)2+(6−4)2
* 4+1+1+0+4
* 10

**3.Agora aplicamos os valores na fórmula de correlação de Pearson:**

* **0.9**

**Neste exemplo podemos observar que há uma forte relação entre X e Y.**

**Limitações do Coeficiente de Correlação de Pearson**

Apesar de sua utilidade, o coeficiente de correlação de Pearson possui algumas limitações:

* **Linearidade**: Pearson mede apenas relações lineares. Ele pode ser enganoso se a relação entre as variáveis for não linear.
* **Sensibilidade a Outliers**: Valores extremos podem influenciar significativamente o valor de r𝑟, distorcendo a verdadeira relação entre as variáveis.
* **Normalidade**: Pearson assume que as variáveis são aproximadamente normalmente distribuídas. Em distribuições não normais, o coeficiente pode não ser um bom indicador da relação entre as variáveis.
* **Independência de Escala**: Pearson não é afetado pela escala das variáveis, mas é sensível à variância conjunta. Assim, pode não detectar correlações em dados escalados ou transformados inadequadamente.

**Métodos de Cálculo**

O coeficiente de correlação de Pearson pode ser calculado usando várias ferramentas estatísticas, incluindo**:**

* **Software Estatístico:** Ferramentas como R, Python (biblioteca pandas), SPSS e SAS oferecem funções integradas para calcular 𝑟**.**
* **Calculadoras Online:** Existem várias calculadoras online que permitem o cálculo rápido do coeficiente de correlação.
* **Planilhas Eletrônicas:** Softwares como Microsoft Excel e Google Sheets possuem funções para calcular a correlação, como CORREL() em Excel.

**Vantagens e Desvantagens**

* **Vantagens**:
  + Simplicidade de cálculo e interpretação.
  + Independente da unidade de medida das variáveis.
* **Desvantagens**:
  + Sensível a valores atípicos.
  + Só mede relações lineares.

**Considerações Finais**

* O coeficiente de correlação de Pearson é uma ferramenta poderosa para analisar a relação linear entre duas variáveis quantitativas. Sua simplicidade e facilidade de interpretação o tornam uma escolha popular em muitas áreas de pesquisa.
* No entanto, os pesquisadores devem estar cientes de suas limitações e considerar outras medidas de correlação ou análise adicional quando necessário. Compreender as nuances do coeficiente de correlação de Pearson e suas aplicações é essencial para realizar análises estatísticas robustas e informadas.

**Resumo Sobre Karl Pearson:**

Karl Pearson foi um renomado matemático e estatístico britânico, nascido em 27 de março de 1857, em Londres, Inglaterra. Ele é amplamente reconhecido como um dos fundadores da estatística moderna e fez contribuições significativas para várias áreas da ciência.

**Contribuições Principais**

* **Coeficiente de Correlação de Pearson**: Karl Pearson desenvolveu o coeficiente de correlação de Pearson, uma medida estatística que quantifica a relação linear entre duas variáveis. [Este coeficiente é amplamente utilizado em diversas disciplinas, incluindo estatística, ciência de dados, economia, psicologia e ciências naturais](https://maestrovirtuale.com/karl-pearson-contribuicoes-para-a-ciencia-e-o-emprego/).
* [**Teste Qui-Quadrado**: Pearson também introduziu o teste qui-quadrado, uma ferramenta estatística usada para testar a independência entre duas variáveis categóricas](https://economiaenegocios.com/o-que-e-o-coeficiente-de-pearson/).
* [**Regressão Linear**: Ele foi pioneiro no desenvolvimento de métodos de regressão linear, que são fundamentais para a análise de relações entre variáveis](https://economiaenegocios.com/o-que-e-o-coeficiente-de-pearson/).

**Vida e Carreira**

Karl Pearson estudou matemática na Universidade de Cambridge e, posteriormente, se interessou por filosofia, literatura e história. Ele foi um dos primeiros a aplicar métodos estatísticos em biologia e psicologia, contribuindo significativamente para a biometria e a psicometria.

**Impacto e Legado**

* [**Fundador da Estatística Moderna**: Pearson é frequentemente considerado o fundador da estatística moderna devido às suas contribuições pioneiras e ao desenvolvimento de técnicas estatísticas avançadas](https://maestrovirtuale.com/karl-pearson-contribuicoes-para-a-ciencia-e-o-emprego/).
* [**Aplicações em Diversas Áreas**: Suas descobertas e métodos são utilizados até hoje em várias áreas do conhecimento, desde a biologia até as ciências sociais e econômicas](https://www.datageeks.com.br/coeficiente-de-correlacao-de-pearson/).
* [**Influência Duradoura**: O trabalho de Pearson continua a influenciar a prática estatística e a pesquisa científica, demonstrando a importância de suas ideias e descobertas para a ciência e a sociedade](https://maestrovirtuale.com/karl-pearson-contribuicoes-para-a-ciencia-e-o-emprego/).

Karl Pearson faleceu em 27 de abril de 1936, mas seu legado perdura, inspirando estatísticos e pesquisadores em todo o mundo.

***Correlação de Kendall ou Coeficiente tau de Kendall***.

A correlação de Kendall, também conhecida como coeficiente tau de Kendall (τ), é uma medida estatística usada para avaliar a associação entre duas variáveis ordinais. [Diferente da correlação de Pearson, que mede a relação linear entre variáveis contínuas, a correlação de Kendall é não-paramétrica e se baseia nos postos das variáveis, não nos seus valores reais](https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_tau_de_Kendall).

Aqui estão alguns pontos-chave sobre a correlação de Kendall:

* **Intervalo de Valores**: O coeficiente τ varia de -1 a +1. Valores próximos a +1 indicam uma forte correlação positiva, enquanto valores próximos a -1 indicam uma forte correlação negativa. [Valores próximos a 0 sugerem pouca ou nenhuma correlação](https://www.blog.psicometriaonline.com.br/o-que-e-correlacao-tau-de-kendall/).
* **Interpretação**: Um τ positivo significa que, à medida que uma variável aumenta, a outra também tende a aumentar. [Um τ negativo indica que, à medida que uma variável aumenta, a outra tende a diminuir](https://www.blog.psicometriaonline.com.br/o-que-e-correlacao-tau-de-kendall/).
* [**Aplicações**: É útil em situações onde os dados não seguem uma distribuição normal ou quando se trabalha com dados ordinais](https://www.blog.psicometriaonline.com.br/o-que-e-correlacao-tau-de-kendall/).

[Dados ordinais são um tipo de dado categórico em que as variáveis têm uma ordem específica, mas as distâncias entre essas categorias não são necessariamente iguais ou conhecidas](https://www.freecodecamp.org/portuguese/news/tipos-de-dados-em-estatistica-tipos-de-dados-nominais-ordinais-intervalares-e-proporcionais-explicados-com-exemplos/). Em outras palavras, os dados ordinais permitem classificar as observações em uma sequência, mas não permitem medir a diferença exata entre elas.

**Exemplos de Dados Ordinais:**

* **Classificações**: Posições em uma corrida (1º lugar, 2º lugar, 3º lugar, etc.).
* **Níveis de Satisfação**: Muito satisfeito, satisfeito, neutro, insatisfeito, muito insatisfeito.
* [**Níveis de Educação**: Ensino fundamental, ensino médio, graduação, pós-graduação](https://www.freecodecamp.org/portuguese/news/tipos-de-dados-em-estatistica-tipos-de-dados-nominais-ordinais-intervalares-e-proporcionais-explicados-com-exemplos/).

[A principal característica dos dados ordinais é que eles indicam uma ordem ou hierarquia, mas não fornecem informações sobre a magnitude das diferenças entre as categorias](https://www.questionpro.com/blog/pt-br/o-que-sao-dados-ordinais/).

As colunas do dataset (*winequality-red.csv*) não se enquadram como dados ordinais. [Elas são exemplos de dados quantitativos contínuos, pois representam medições numéricas que podem assumir qualquer valor dentro de um intervalo contínuo](https://www.freecodecamp.org/portuguese/news/tipos-de-dados-em-estatistica-tipos-de-dados-nominais-ordinais-intervalares-e-proporcionais-explicados-com-exemplos/).

**Tipos de Dados:**

* **Dados Ordinais**: Representam categorias com uma ordem específica, mas sem uma diferença mensurável entre elas (por exemplo, níveis de satisfação: satisfeito, neutro, insatisfeito).
* **Dados Quantitativos Contínuos**: Representam medições numéricas que podem assumir qualquer valor dentro de um intervalo contínuo (por exemplo, pH, densidade, teor de álcool).

[As colunas da tabela de vinhos tintos, como “Alcohol”, “Volatile Acidity”, “pH”, etc., são medições contínuas e não categóricas, portanto, não são dados ordinais](https://www.freecodecamp.org/portuguese/news/tipos-de-dados-em-estatistica-tipos-de-dados-nominais-ordinais-intervalares-e-proporcionais-explicados-com-exemplos/).

**Fórmula Básica de Kendall**

A fórmula básica para calcular o coeficiente de Kendall é:

τ=​

Onde:

* **C** é o número de pares concordantes.
* **D** é o número de pares discordantes.

**Propriedades**

* O [denominador](https://pt.wikipedia.org/wiki/Denominador) é o número total de combinações de pares, então, o coeficiente deve estar no intervalo −1≤𝜏≤1-1 ≤ τ ≤ 1.
* Se a concordância entre as duas classificações for perfeita (isto é, se as duas classificações forem iguais), o coeficiente tem valor 1.
* Se a discordância entre as duas classificações for perfeita (isto é, se uma classificação for o reverso da outra), o coeficiente tem valor -1.
* Se 𝑋**C** e 𝑌**D** forem [independentes](https://pt.wikipedia.org/wiki/Independ%C3%AAncia_(estat%C3%ADstica)), espera-se que o coeficiente seja próximo de zero.

**Pares Concordantes e Discordantes:**

* **Pares Concordantes**: Um par de observações ((x\_i, y\_i)) e ((x\_j, y\_j)) é concordante se a ordem de (x\_i) e (x\_j) concorda com a ordem de (y\_i) e (y\_j). Ou seja, se (x\_i < x\_j) e (y\_i < y\_j) ou (x\_i > x\_j) e (y\_i > y\_j).
* **Pares Discordantes**: Um par de observações ((x\_i, y\_i)) e ((x\_j, y\_j)) é discordante se a ordem de (x\_i) e (x\_j) discorda da ordem de (y\_i) e (y\_j). Ou seja, se (x\_i < x\_j) e (y\_i > y\_j) ou (x\_i > x\_j) e (y\_i < y\_j).

**Fórmula para Dados com Empates:**

Quando há empates nos dados, a fórmula de Kendall pode ser ajustada para levar isso em consideração. A fórmula ajustada é:

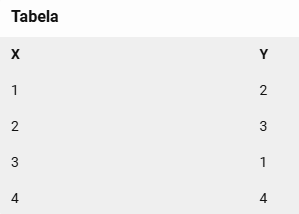
τ=(C+D+Tx​)(C+D+Ty​)​C−D​

**Onde:**

* **T\_x** é o número de pares empatados em (x).
* **T\_y** é o número de pares empatados em (y).

Exemplo de Cálculo

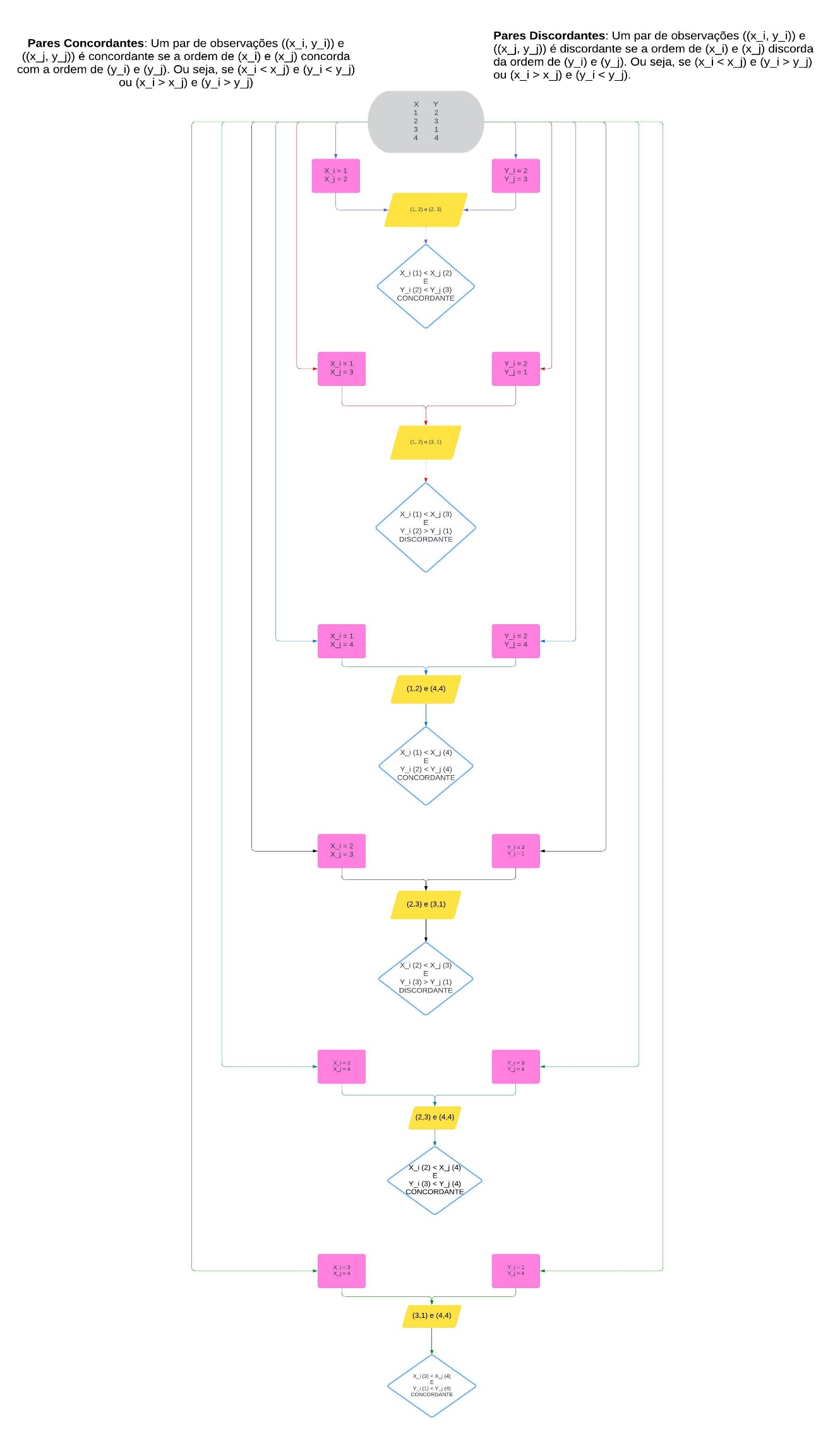
Vamos supor que temos as seguintes observações para duas variáveis (X) e (Y):



1. Contar Pares Concordantes ©:
   * Pares (1,2) e (2,3): Concordante
   * Pares (1,2) e (3,1): Discordante
   * Pares (1,2) e (4,4): Concordante
   * Pares (2,3) e (3,1): Discordante
   * Pares (2,3) e (4,4): Concordante
   * Pares : Concordante

Total de pares concordantes (C = 4).

1. Contar Pares Discordantes (D):
   * Total de pares discordantes (D = 2).



1. **Cálculo do Coeficiente de Kendall**

τ=​ = = ​ = ​≈ 0.333

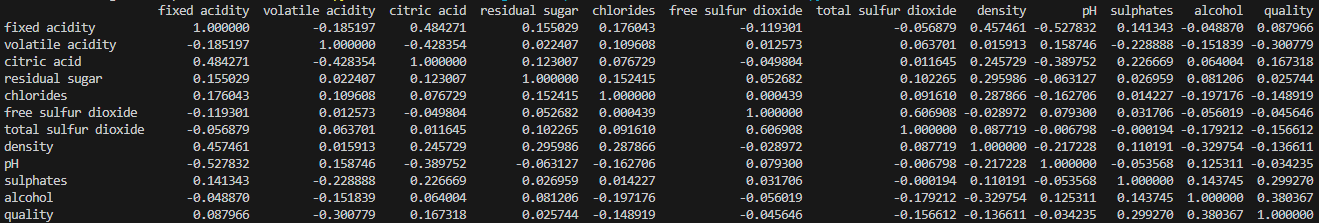
Portanto, o coeficiente de Kendall (τ) é aproximadamente 0.333, indicando uma correlação positiva moderada entre as variáveis (X) e (Y).

**Demonstração de aplicação do Coeficiente Tau de Kendall**.

O código python abaixo recebe a variável data(*winequality-red.csv*) onde esta sendo aplicado o método corr(), onde recebe o coeficiente tau de Kendall τ.

**Saída da Correlação de Kendall**

A matriz de correlação de Kendall resultante mostra os coeficientes de correlação entre cada par de variáveis. Aqui estão alguns pontos importantes da análise:

1. **Correlação Positiva Forte**:
   * *fixed acidity***e***citric acid*: 0.484 - Indica que, à medida que a acidez fixa aumenta, a acidez cítrica também tende a aumentar.
   * *free sulfur dioxide***e***total sulfur dioxide*: 0.607 - Sugere que há uma forte relação positiva entre o dióxido de enxofre livre e total.
2. **Correlação Negativa**:
   * *volatile acidity***e***citric acid*: -0.428 - Indica que, à medida que a acidez volátil aumenta, a acidez cítrica tende a diminuir.
   * *pH***e***fixed acidity*: -0.528 - Mostra que, à medida que a acidez fixa aumenta, o pH tende a diminuir.
3. **Correlação com *quality***:
   * *alcohol***e***quality*: 0.380 - Sugere que há uma correlação positiva moderada entre o teor de álcool e a qualidade do vinho.
   * *volatile acidity***e***quality*: -0.301 - Indica uma correlação negativa, sugerindo que vinhos com maior acidez volátil tendem a ter menor qualidade.

**Interpretação**

Embora a correlação de Kendall seja mais recomendada para dados ordinais, ela ainda pode ser útil para dados contínuos, especialmente quando a relação entre as variáveis não é linear. A correlação de Kendall mede a força e a direção de uma relação monotônica entre duas variáveis, o que pode ser mais robusto em presença de outliers ou distribuições não normais.

**Conclusão**

A matriz de correlação de Kendall fornece insights valiosos sobre as relações entre as variáveis químicas e a qualidade do vinho. Mesmo que as variáveis não sejam ordinais, a correlação de Kendall pode capturar relações monotônicas que outras medidas, como a correlação de Pearson, podem não detectar.