# DATA SCIENCE

## AULA 10 – SISTEMA DE RECOMENDAÇÃO I

PROFª. ANA CAROLINA B. ALBERTON

# INTRODUÇÃO

Variáveis

- Existe uma relação matemática entre estas duas variáveis?
- Se existe, como posso medir sua força?
- Poderia usar essa relação para fazer previsões?

No olhômetro, não conseguimos identificar uma relação, principalmente em um conjunto de dados maior.

| Idade | Custo |
|---|---|
| 18 | 871 |
| 23 | 1132 |
| 28 | 1242 |
| 33 | 1356 |
| 38 | 1488 |
| 43 | 1638 |
| 48 | 2130 |
| 53 | 2454 |
| 58 | 3066 |
| 63 | 4090 |

Como então podemos criar essa relação, imaginando um conjunto de dados com mais de 5000 linhas?

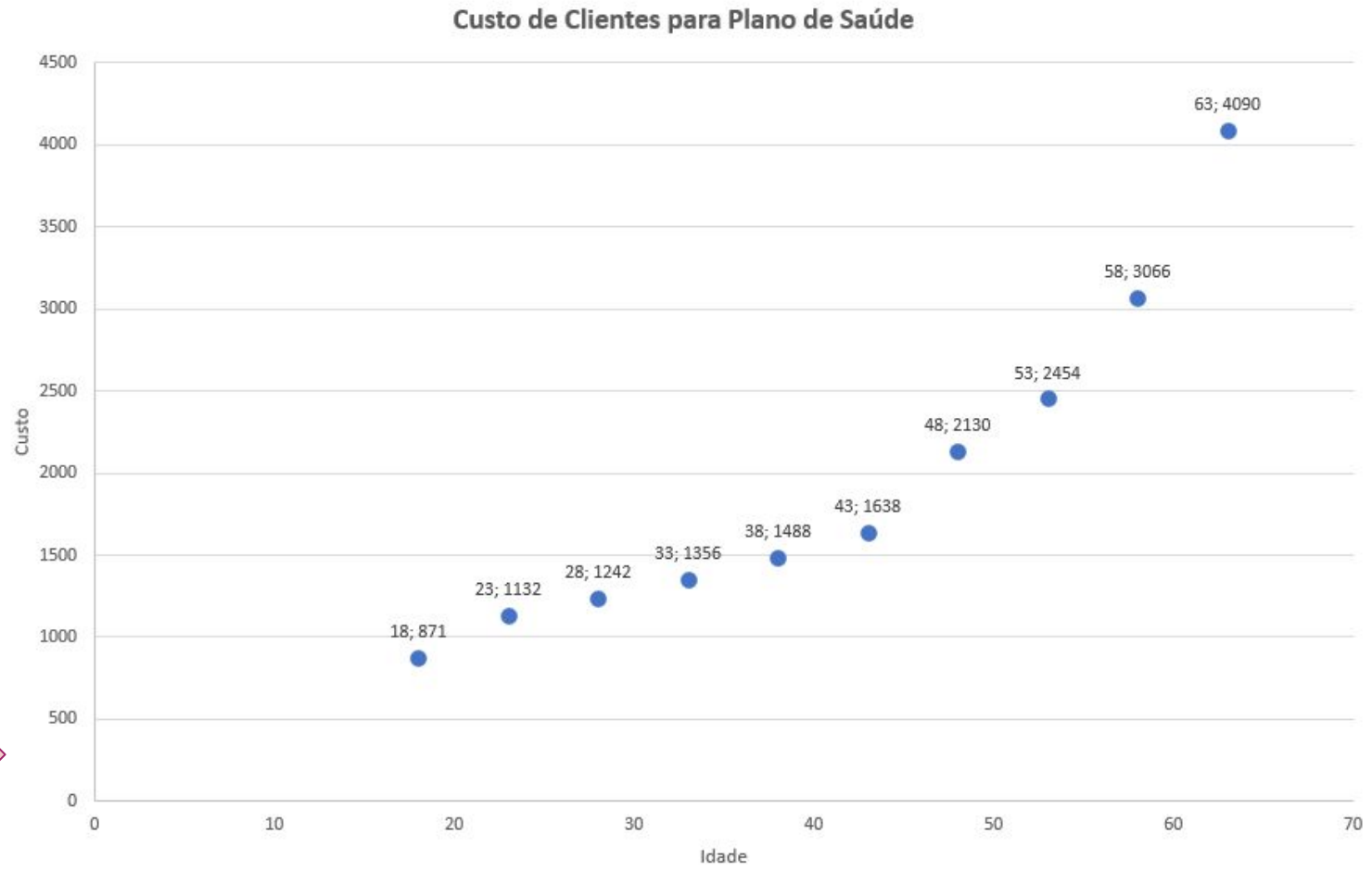
Essa é a questão a ser resolvida na aula de hoje.

Vamos então seguir uma linha de raciocínio… me acompanhem (mesmo que de forma remota)…

# GRÁFICO DE DISPERSÃO

| Idade | Custo |
|---|---|
| 18 | 871 |
| 23 | 1132 |
| 28 | 1242 |
| 33 | 1356 |
| 38 | 1488 |
| 43 | 1638 |
| 48 | 2130 |
| 53 | 2454 |
| 58 | 3066 |
| 63 | 4090 |

Vamos plotar essas variáveis para visualizar melhor

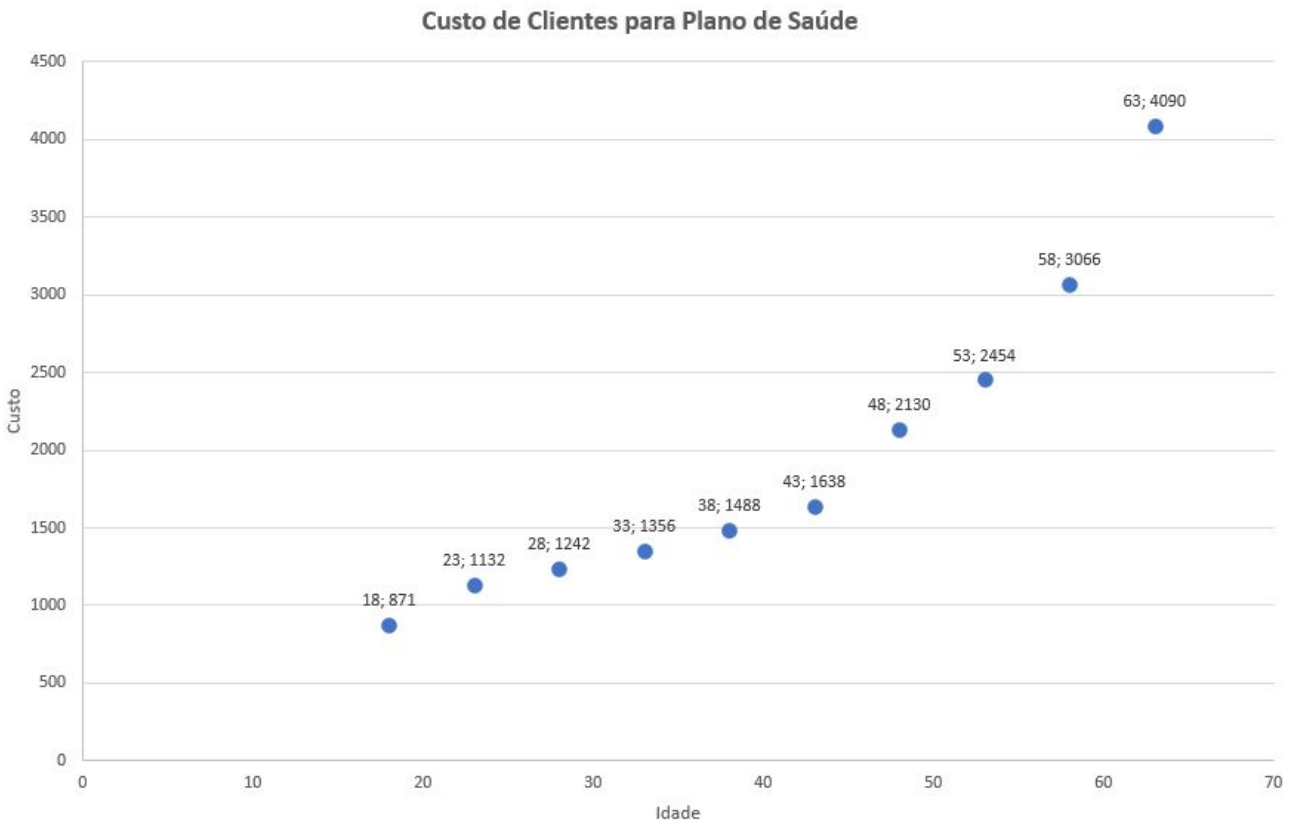


# GRÁFICO DE DISPERSÃO

| Idade | Custo |
|---|---|
| 18 | 871 |
| 23 | 1132 |
| 28 | 1242 |
| 33 | 1356 |
| 38 | 1488 |
| 43 | 1638 |
| 48 | 2130 |
| 53 | 2454 |
| 58 | 3066 |
| 63 | 4090 |

É importante ter bem claro essas definições em vermelho

Eixo Y (Vertical)
Variável de Resposta
Ou Dependente
Na regressão é o que queremos Prever




Olhando assim, Vcs já conseguem me dizer qual vai ser o custo para o plano de saúde de um paciente com 45 anos de idade?

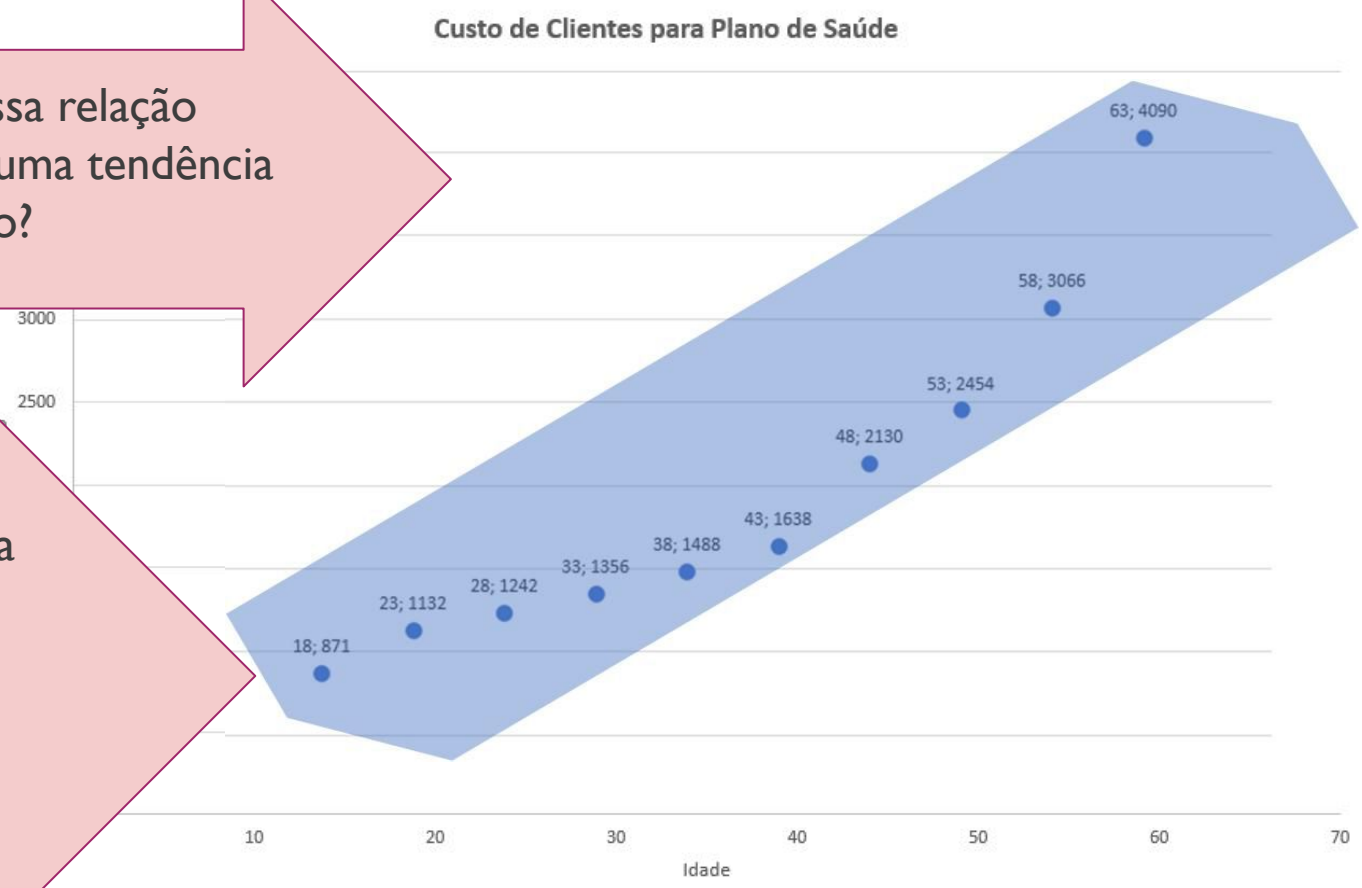
Ainda não né! Vamos ver o resto

Eixo X (Horizontal)
Variável Explanatória ou Independente
Na regressão é o que explica,
ou usamos para prever

# REGRESSÃO LINEAR

Como podemos ver, essa relação entre as variáveis tem uma tendência a ser uma linha, correto?

Não estou dizendo que é uma linha, apenas que possui uma tendência. Poderia ser uma curva exponencial, uma parábola…. mas é uma linha



# CORRELAÇÃO (r)

Existe um valor que diz para a gente qual a relação entre elas, chamado de correlação

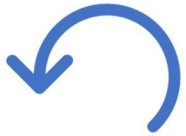
- Mostra a força e a direção da relação entre variáveis
  - Pode ser um valor entre -1 e 1
  - A correlação de A ~ B é a mesma que B ~ A

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

# CORRELAÇÃO (R)

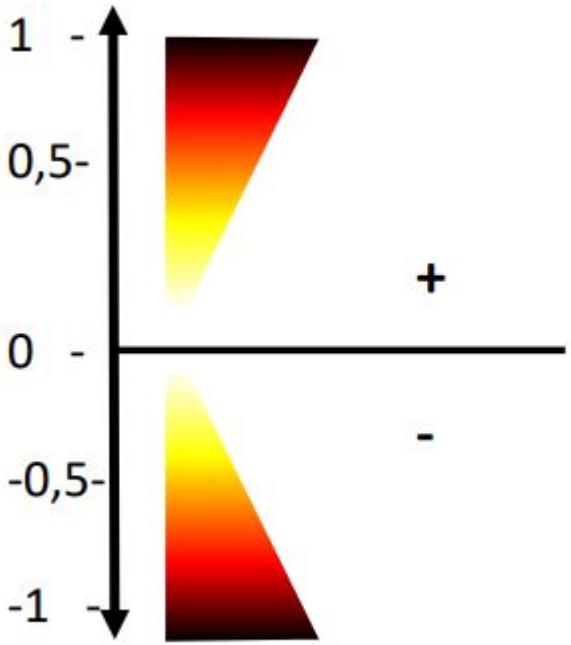
Mostra a força e a direção da relação entre variáveis

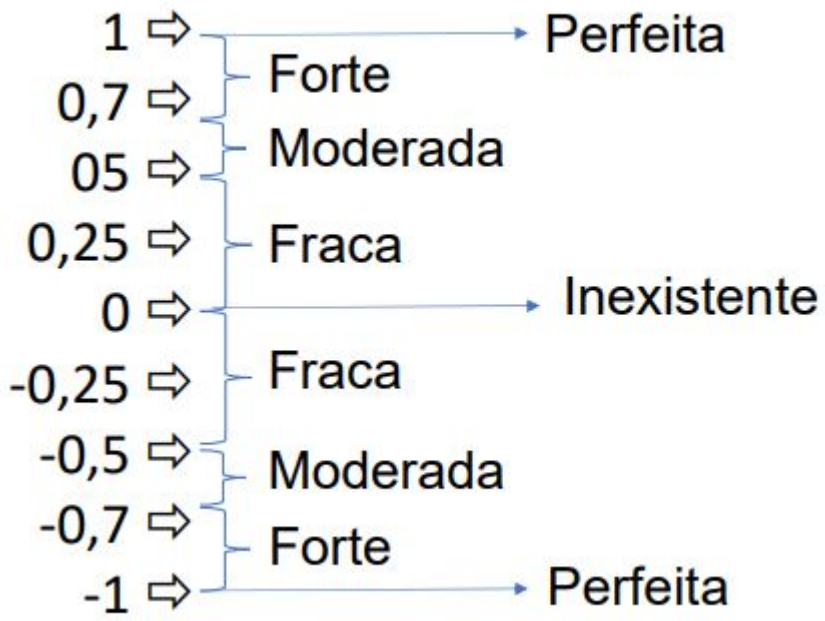

Pode ser um valor entre -1 e 1



A correlação de A ~ B é a mesma que B ~ A

# CORRELAÇÃO - FORÇA E DIREÇÃO



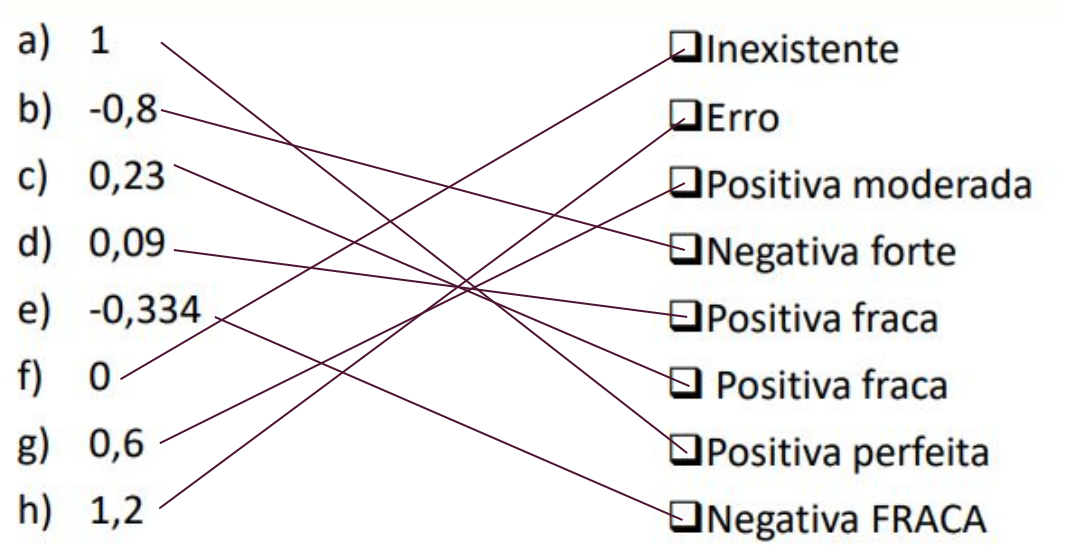
# EXEMPLO

Relacione:

| | | |
|---|---|---|
| a) | 1 | ❑Inexistente |
| b) | -0,8 | ❑Erro |
| c) | 0,23 | ❑Positiva moderada |
| d) | 0,09 | ❑Negativa forte |
| e) | -0,334 | ❑Positiva fraca |
| f) | 0 | ❑ Positiva fraca |
| g) | 0,6 | ❑Positiva perfeita |
| h) | 1,2 | ❑Negativa FRACA |

# EXEMPLO - resposta

Relacione:



# CORRELAÇÃO

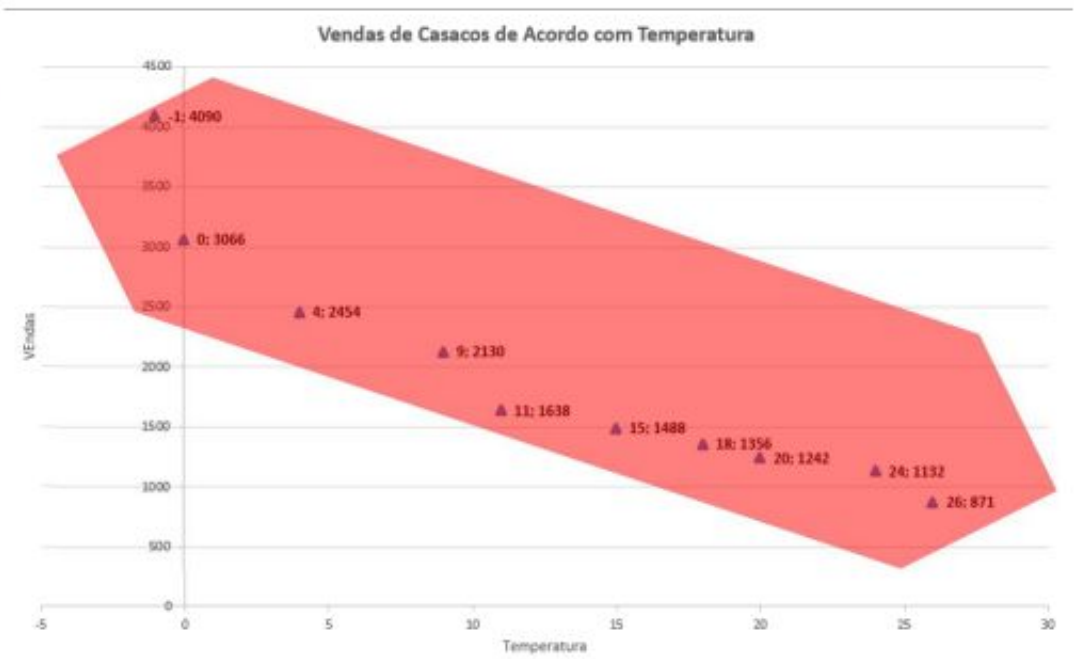
## Forte - Fraca


Cor: 0,93092


Cor: -0,22765

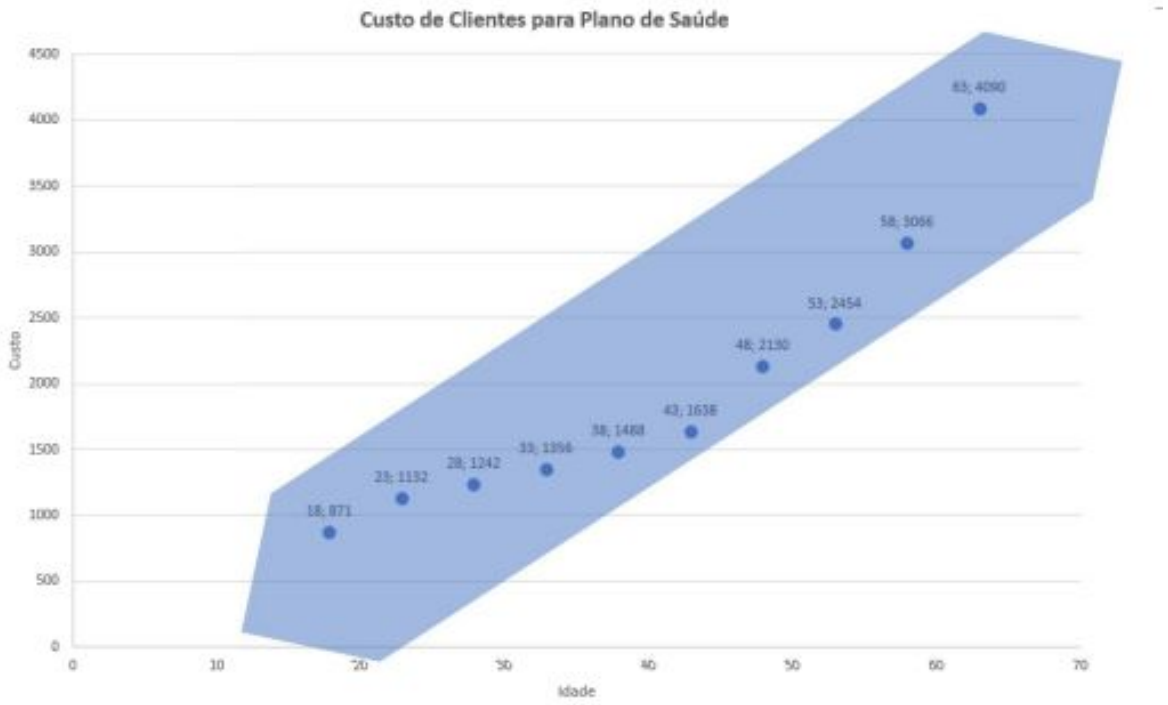
# Positiva - Negativa


Cor: 0,93092



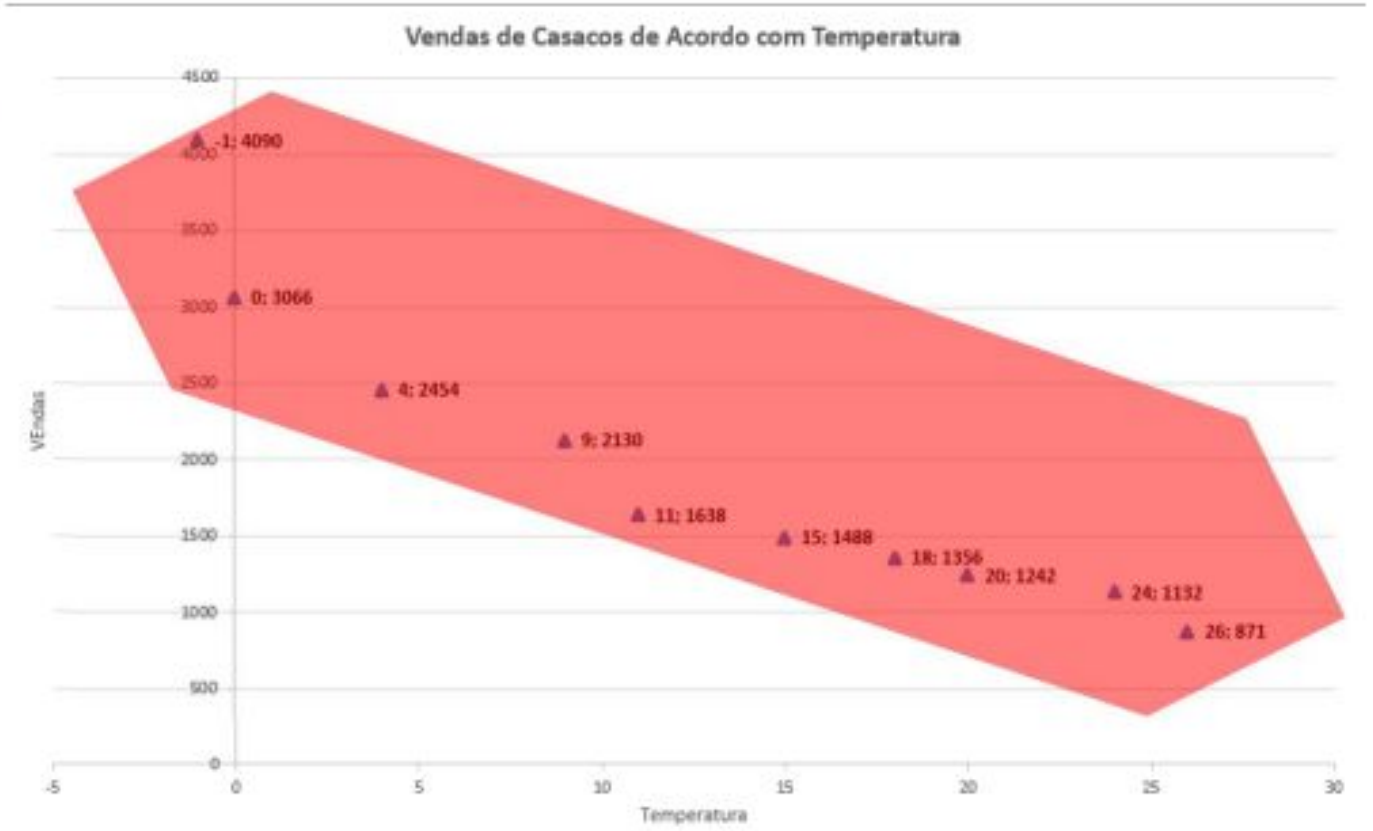

Cor: -0,93092

# CORRELAÇÃO POSITIVA




| Idade | Custo |
|---|---|
| 18 | 871 |
| 23 | 1132 |
| 28 | 1242 |
| 33 | 1356 |
| 38 | 1488 |
| 43 | 1638 |
| 48 | 2130 |
| 53 | 2454 |
| 58 | 3066 |
| 63 | 4090 |

# CORRELAÇÃO NEGATIVA




| Temperatura | Vendas |
|---|---|
| -1 | 4090 |
| 0 | 3066 |
| 4 | 2454 |
| 9 | 2130 |
| 11 | 1638 |
| 15 | 1488 |
| 18 | 1356 |
| 20 | 1242 |
| 24 | 1132 |
| 26 | 871 |

# COEFICIENTE DE DETERMINAÇÃO ($R^2$)

Temos também um outro valor, derivado do R.
O coeficiente de determinação, que é o R ao quadrado

- Mostra o quanto o modelo consegue explicar os valores
- Quanto maior, mais explicativo ele é
- O restante da variabilidade está em variáveis não incluídas no modelo
- Varia entre zero até 1 (Sempre positivo)
- Calcula-se com o quadrado do coeficiente de correlação (R)

# COEFICIENTE DE DETERMINAÇÃO ($R^2$)

- Correlação = 0,93
- $R^2$=0,86
- 86% da variável dependente consegue ser explicada pelas variáveis explanatórias presentes no modelo

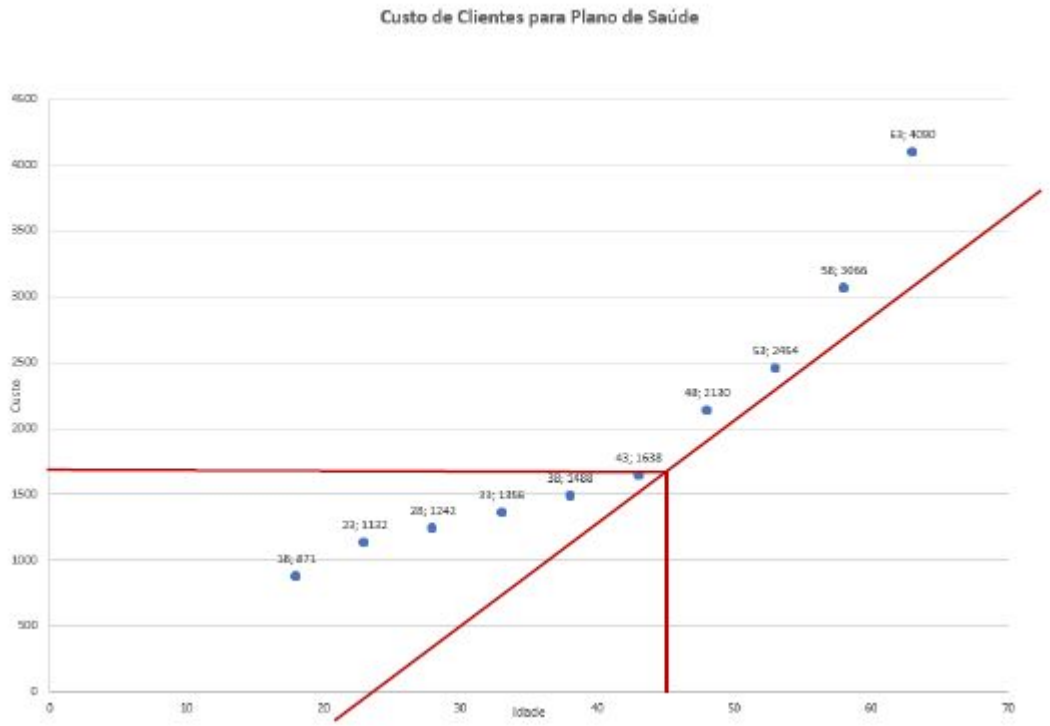
Esse é o significado do valor dele.
Muito interessante, por sinal!

Então quando temos uma boa correlação e um bom $R^2$, podemos aplicar a função de regressão linear

| Idade | Custo |
|---|---|
| 18 | 871 |
| 23 | 1132 |
| 28 | 1242 |
| 33 | 1356 |
| 38 | 1488 |
| 43 | 1638 |
| 48 | 2130 |
| 53 | 2454 |
| 58 | 3066 |
| 63 | 4090 |

# REGRESSÃO LINEAR - PREVISÃO

- Previsão: Qual vai ser o custo de um cliente com 45 anos de idade?




O correto para criar essa previsão é encontrar aquela linha vermelha, ou seja, encontrar a equação da reta Y=ax+b (lembram disso?)

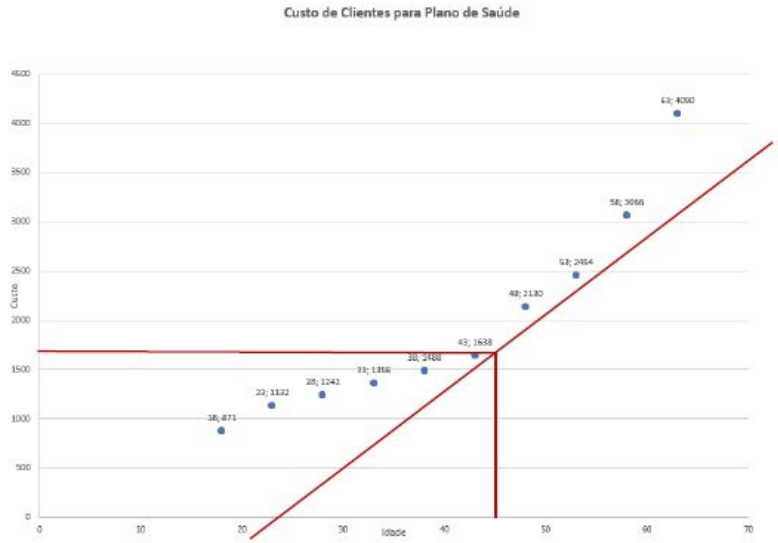
# REGRESSÃO LINEAR - PREVISÃO

- Como a linha é construída?
  Interseção: Ponto de Encontro da Linha no Eixo Y : X=0
- Inclinação: a cada unidade que aumenta a variável Independente (x),

  a variável de resposta (y) sobe o valor da inclinação

  y = ax + b

  y = inclinação*x + Interseção



# REGRESSÃO LINEAR - PREVISÃO

Se pegarmos usando os dados, seria só somar.
Mas então porque essa conta não bate, se tentar encontrar a equação da reta?
Porque não consideramos os erros

Exemplo para encontrar o custo para idade = 34:

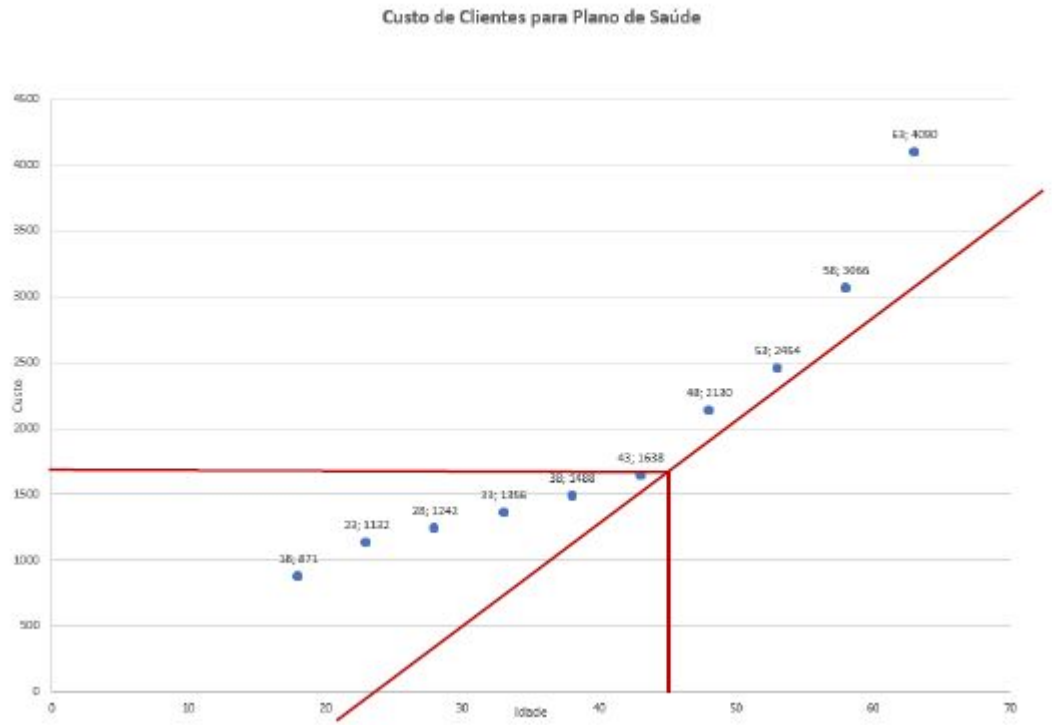
Intersecção: -558,94

Inclinação: 61,86

Previsão:

33 anos: 1356

34 anos: 1356 + 61,86 = 1417,86



# REGRESSÃO LINEAR - PREVISÃO

- Quando analisamos dados que sugerem a existência de uma relação funcional entre duas variáveis, surge então o problema de se determinar uma função matemática que exprime esse relacionamento, ou seja, uma equação de regressão.
- Ao imaginar uma relação funcional entre duas variáveis, digamos X e Y, estamos interessados numa função que explique grande parte da variação de Y por X. Entretanto, uma parcela da variabilidade de Y não explicada por X será atribuída ao acaso, ou seja, ao erro aleatório.
- Quando se estuda a variação de uma variável Y em função de uma variável X, dizemos que Y é a variável dependente e que X é a variável explanatória (ou independente)

# REGRESSÃO LINEAR SIMPLES

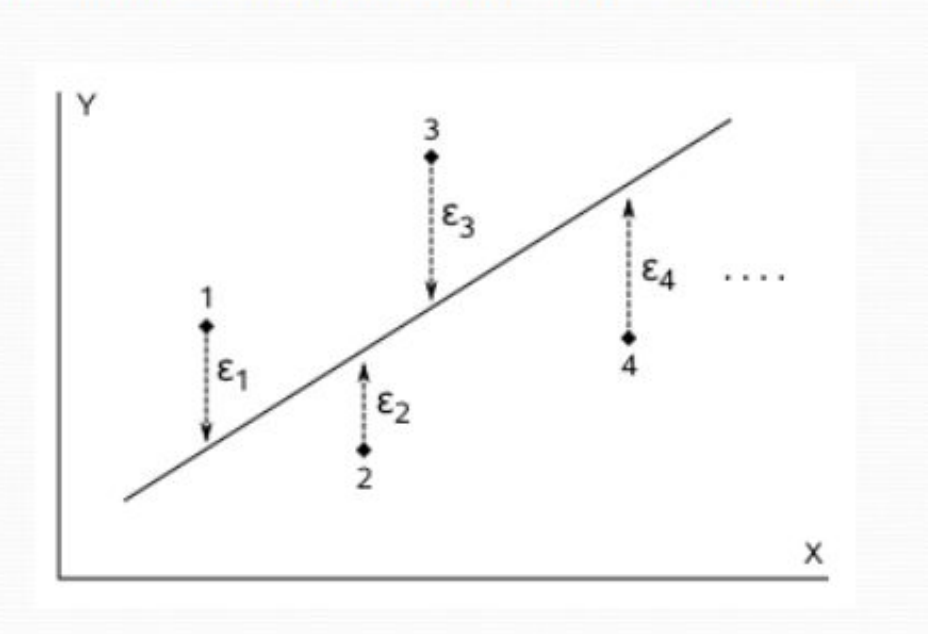
- Formalmente, a análise de regressão parte de um conjunto de observações pareadas,(x1,y1),(x2,y2).. e considere que podemos escrever a relação da seguinte maneira:

$$y_i = \alpha + \beta x_i + E$$

- α e β devem ser estimados
- E é  erro aleatório para  a i-ésima observação

# ESTIMAÇÃO DE PARÂMETROS

- O objetivo é estimar valores para $\alpha$ e $\beta$ através dos dados fornecidos pela amostra
- Além disso, encontrar a reta que passe o mais próximo possível dos pontos observados segundo um critério estabelecido



# MÉTODO DOS MÍNIMOS QUADRADOS

- É usado para estimar os parâmetros do modelo e consiste em fazer com que a soma dos erros quadráticos seja menor possível ou seja este método consiste em obter os valores de **α e β** que minimizam a expressão

$$S = \sum \varepsilon_i = \sum (Y_i - \alpha - \beta x_i)^2$$

- Aplicando-se as derivadas parciais à expressão acima, e igualando-se a zero, acharemos as estimativas para **α e β**

$$a = \frac{\sum y_i - b \sum x_i}{n}$$

e

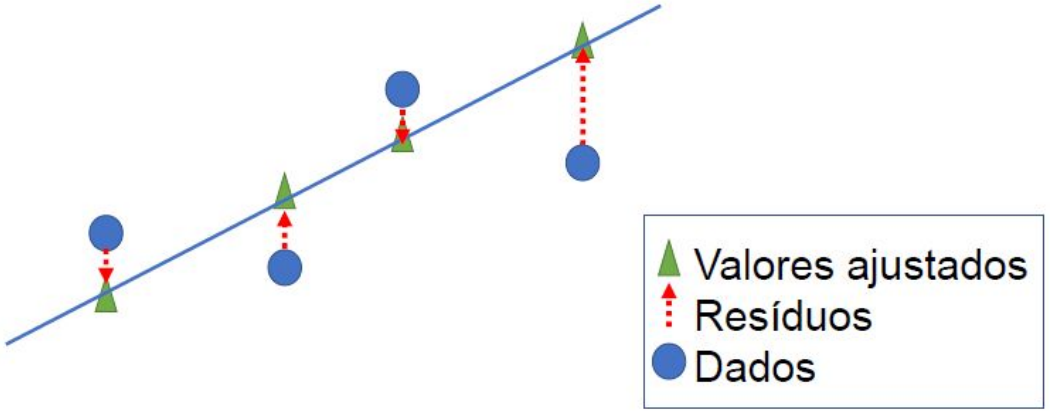
$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

# RESÍDUOS

- A diferença entre os valores observados e os preditos será chamada de **resíduos**

$$e_i = y_i - \hat{y}_i$$

- O resíduo relativo é a i-ésima observação e pode ser chamada de erro aleatório como mostra abaixo

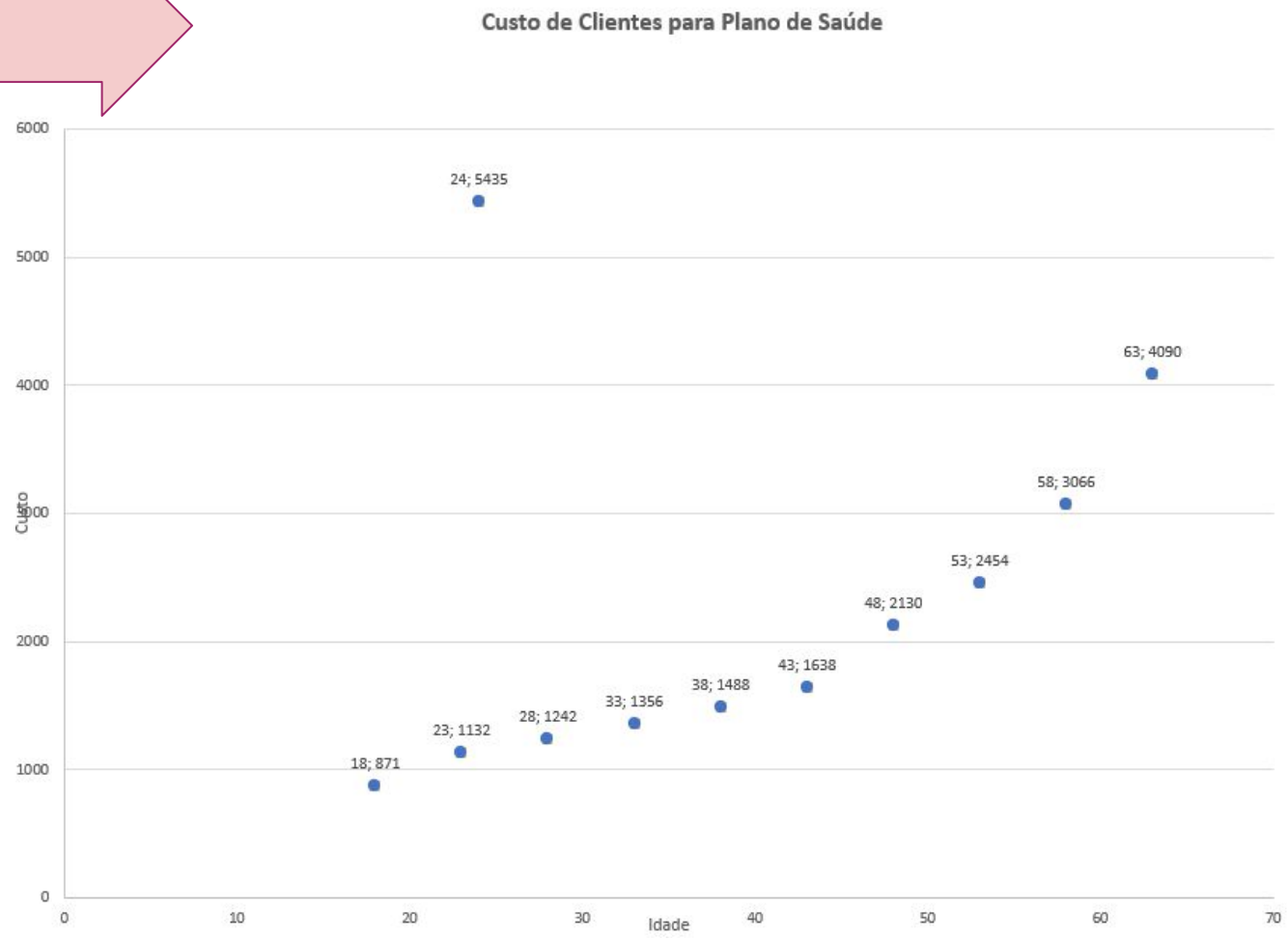


# OUTLIERS

Nesses casos os outiliers interferem muito, por isso é sempre bom avaliar

Por exemplo aqui. acrescentei essa terceira linha, olhem como ficou a minha correlação

| Idade | Custo |
|---|---|
| 18 | 871 |
| 23 | 1132 |
| 24 | 5435 |
| 28 | 1242 |
| 33 | 1356 |
| 38 | 1488 |
| 43 | 1638 |
| 48 | 2130 |
| 53 | 2454 |
| 58 | 3066 |
| 63 | 4090 |




Antiga correlação= 0,93

Nova correlação= 0,34

# Desafio -

Usando o arquivo .ipynb com os códigos em python e esse conjunto de dados, responda:

1) Qual a correlação?
2) Qual o coeficiente de determinação?
3) Essa correlação é forte, moderada ou fraca? positiva ou negativa? Podemos explorar a regressão linear?
4) Qual seria a quantidade de gols feitos para um jogador que ficou em 23° lugar? (dica: use os valores gerados dos coeficientes de intercessão e de inclinação)

| Posição | Gols Feitos |
|---|---|
| 1º | 86 |
| 2º | 60 |
| 3º | 61 |
| 4º | 64 |
| 5º | 51 |
| 6º | 39 |
| 7º | 44 |
| 8º | 42 |
| 9º | 50 |
| 10º | 46 |
| 11º | 44 |
| 12º | 39 |
| 13º | 44 |
| 14º | 38 |
| 15º | 31 |
| 16º | 36 |
| 17º | 27 |
| 18º | 24 |
| 19º | 31 |
| 20º | 18 |