
DATA SCIENCE

AULA 10 – SISTEMA DE RECOMENDAÇÃO I

PROF^a. 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?

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

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

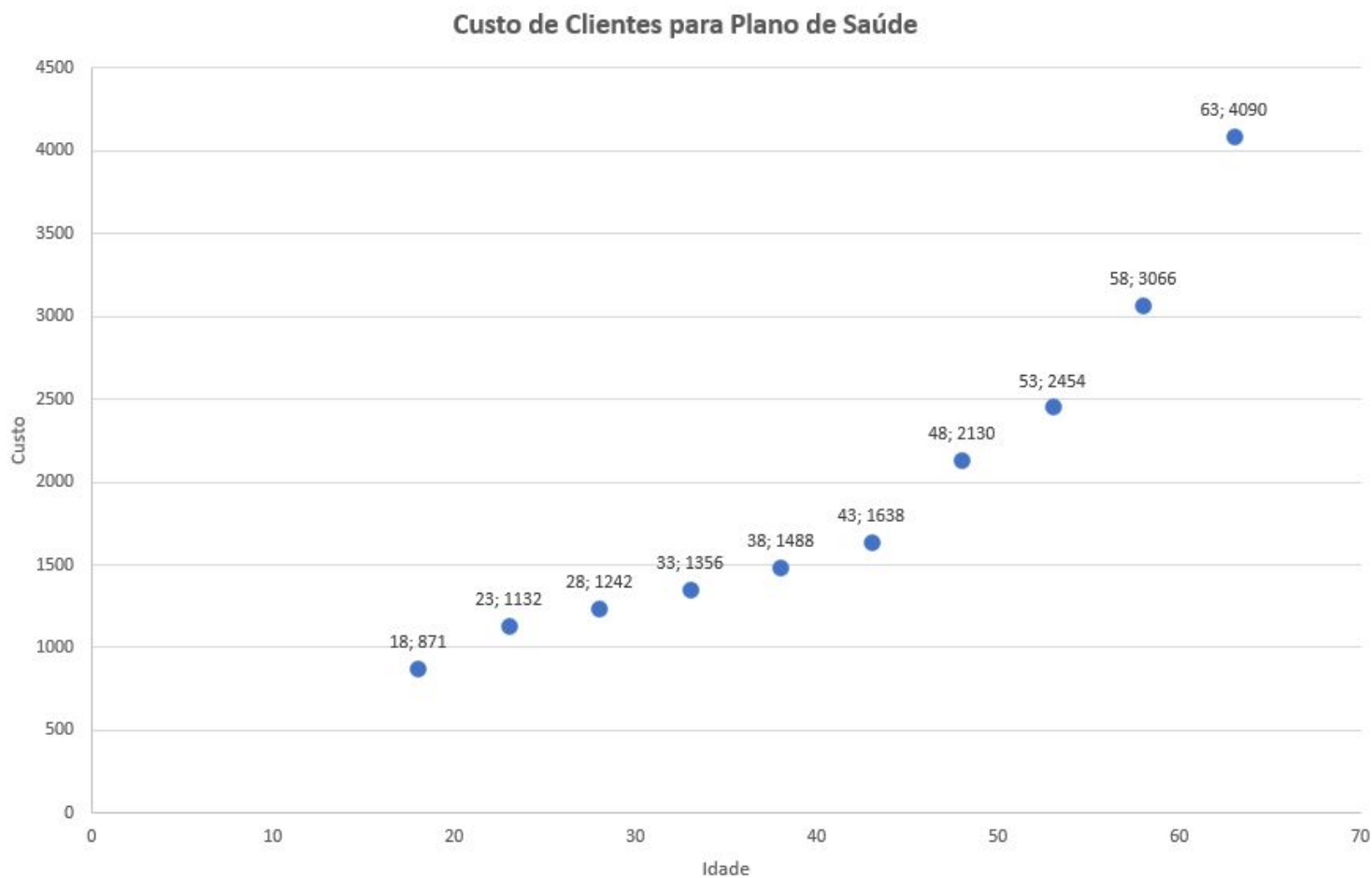
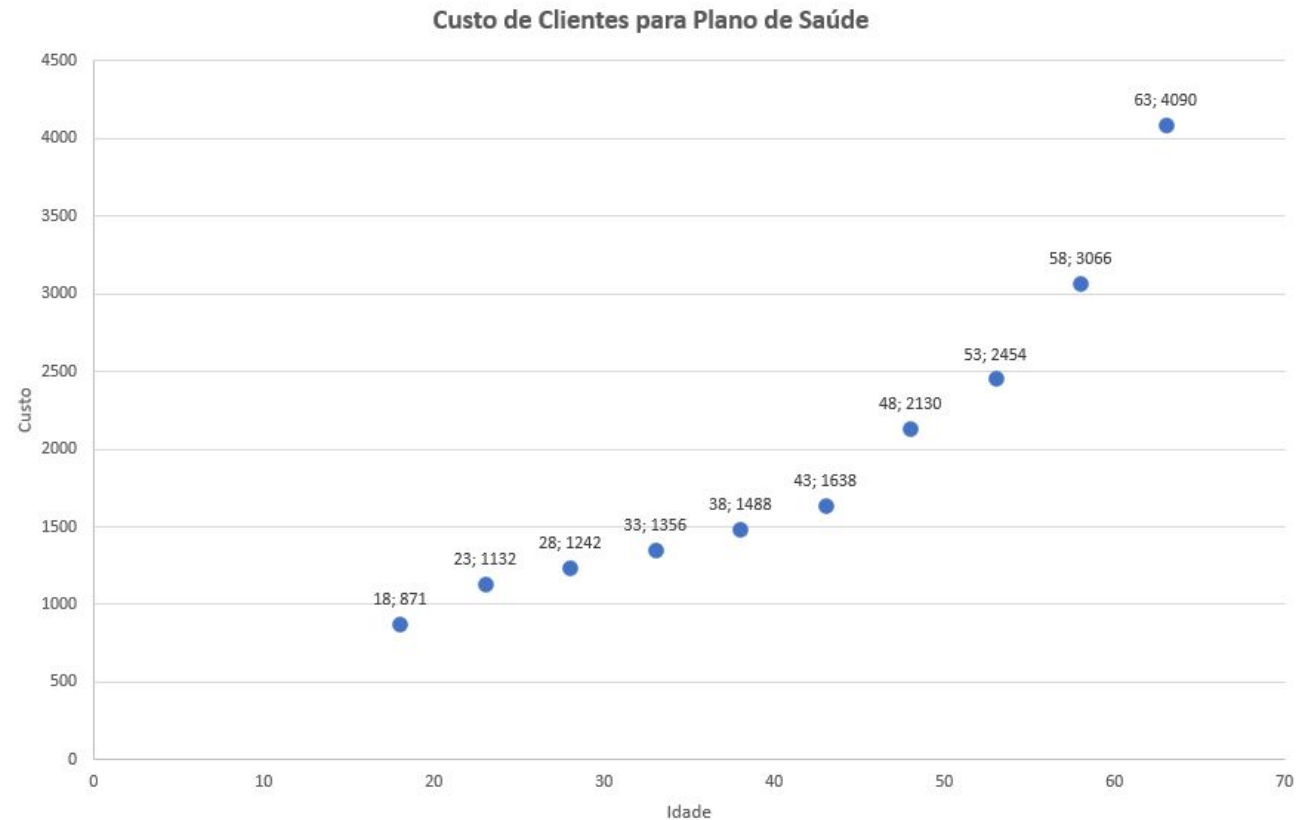


GRÁFICO DE DISPERSÃO

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

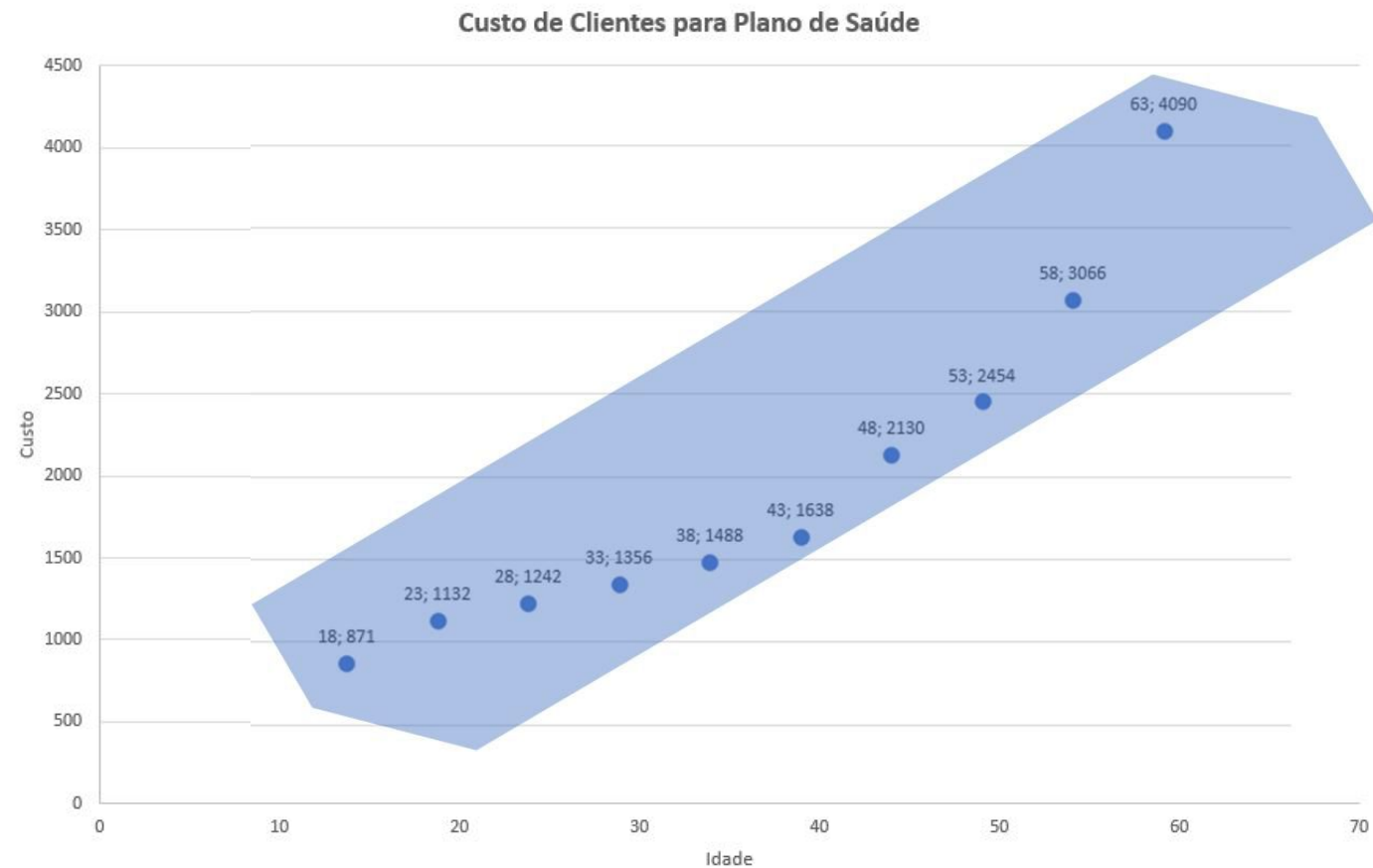
Qual vai ser o custo para o plano de saúde de um paciente com 45 anos de idade?

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



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

REGRESSÃO LINEAR



CORRELAÇÃO (R)

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

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

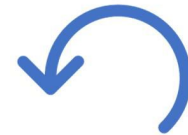
CORRELAÇÃO (R)



Mostra a força e a direção da relação entre variáveis

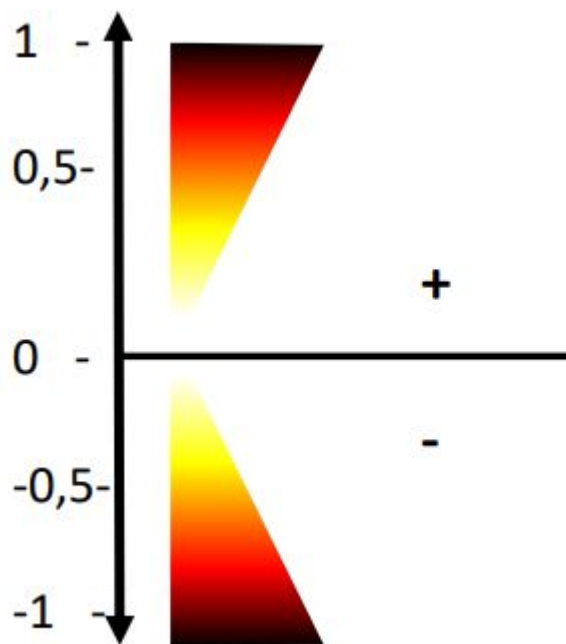


Pode ser um valor entre -1 e 1



A correlação de $A \sim B$ é a mesma que $B \sim A$

CORRELAÇÃO - FORÇA E DIREÇÃO



1	⇒	Perfeita
0,7	⇒	Forte
0,5	⇒	Moderada
0,25	⇒	Fraca
0	⇒	Inexistente
-0,25	⇒	Fraca
-0,5	⇒	Moderada
-0,7	⇒	Forte
-1	⇒	Perfeita

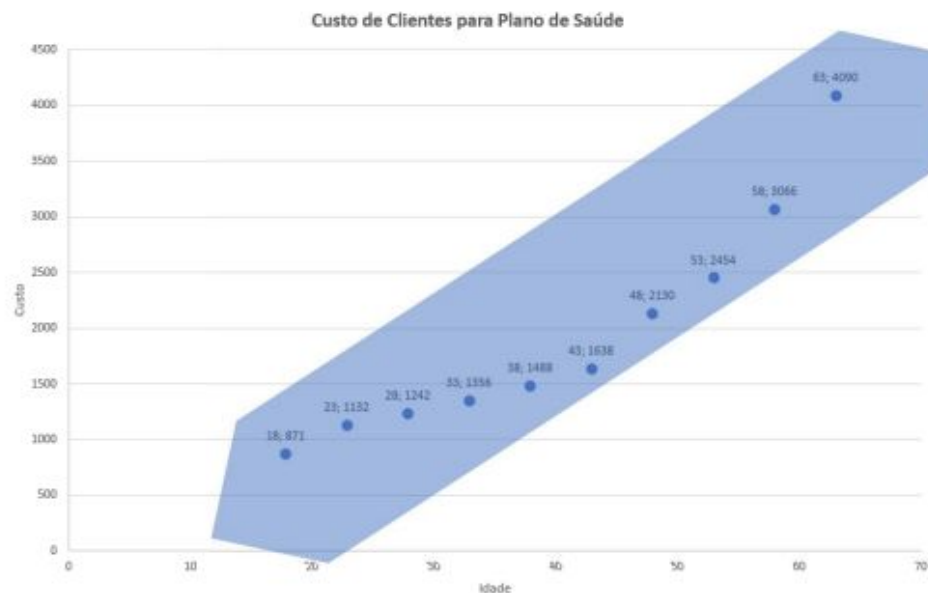
EXEMPLO

Relaçõe:

- | | |
|-----------|--------------------------------------------|
| a) 1 | <input type="checkbox"/> Inexistente |
| b) -0,8 | <input type="checkbox"/> Erro |
| c) 0,23 | <input type="checkbox"/> Positiva moderada |
| d) 0,09 | <input type="checkbox"/> Negativa forte |
| e) -0,334 | <input type="checkbox"/> Positiva fraca |
| f) 0 | <input type="checkbox"/> Positiva fraca |
| g) 0,6 | <input type="checkbox"/> Positiva perfeita |
| h) 1,2 | <input type="checkbox"/> Negativa FRACA |

CORRELAÇÃO

Forte - Fraca

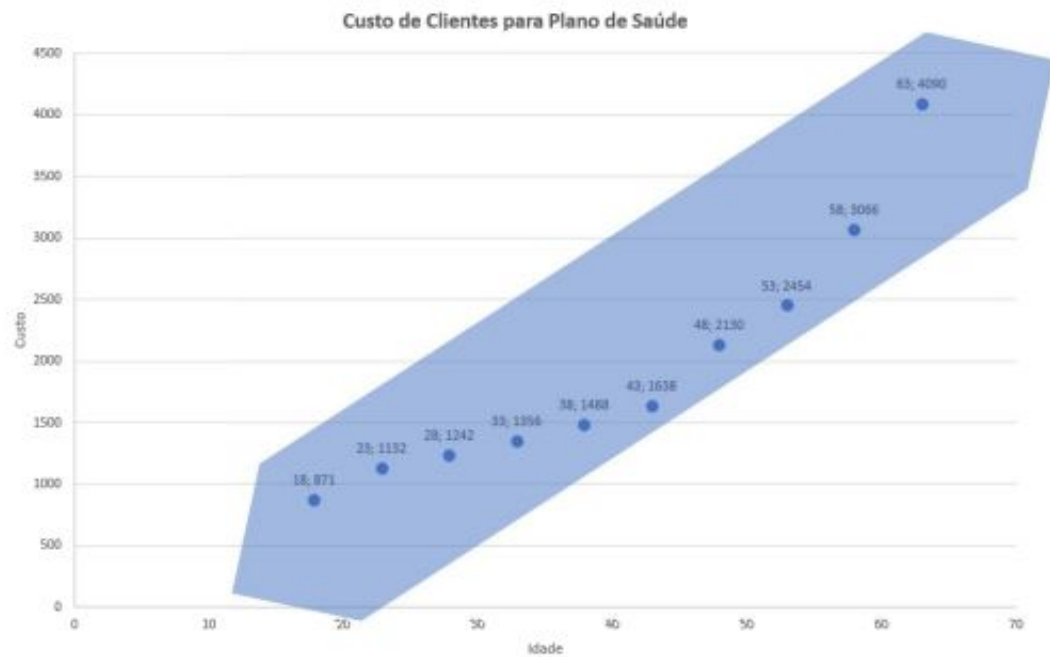


Cor: 0,93092

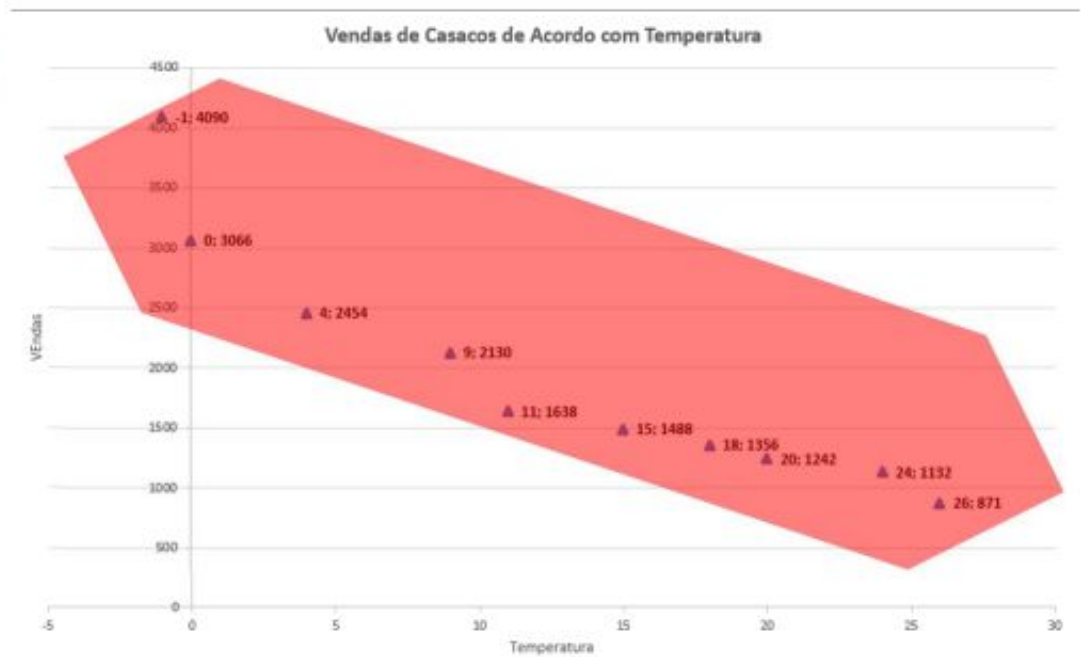


Cor: -0,22765

Positiva - Negativa

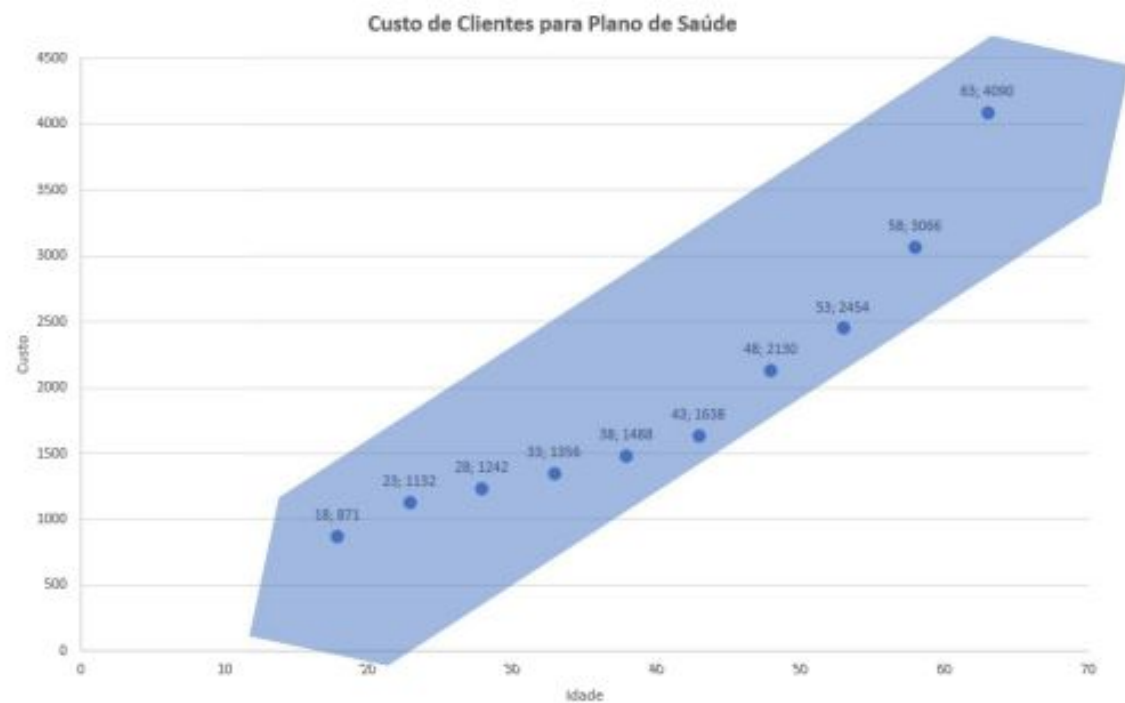


Cor: 0,93092

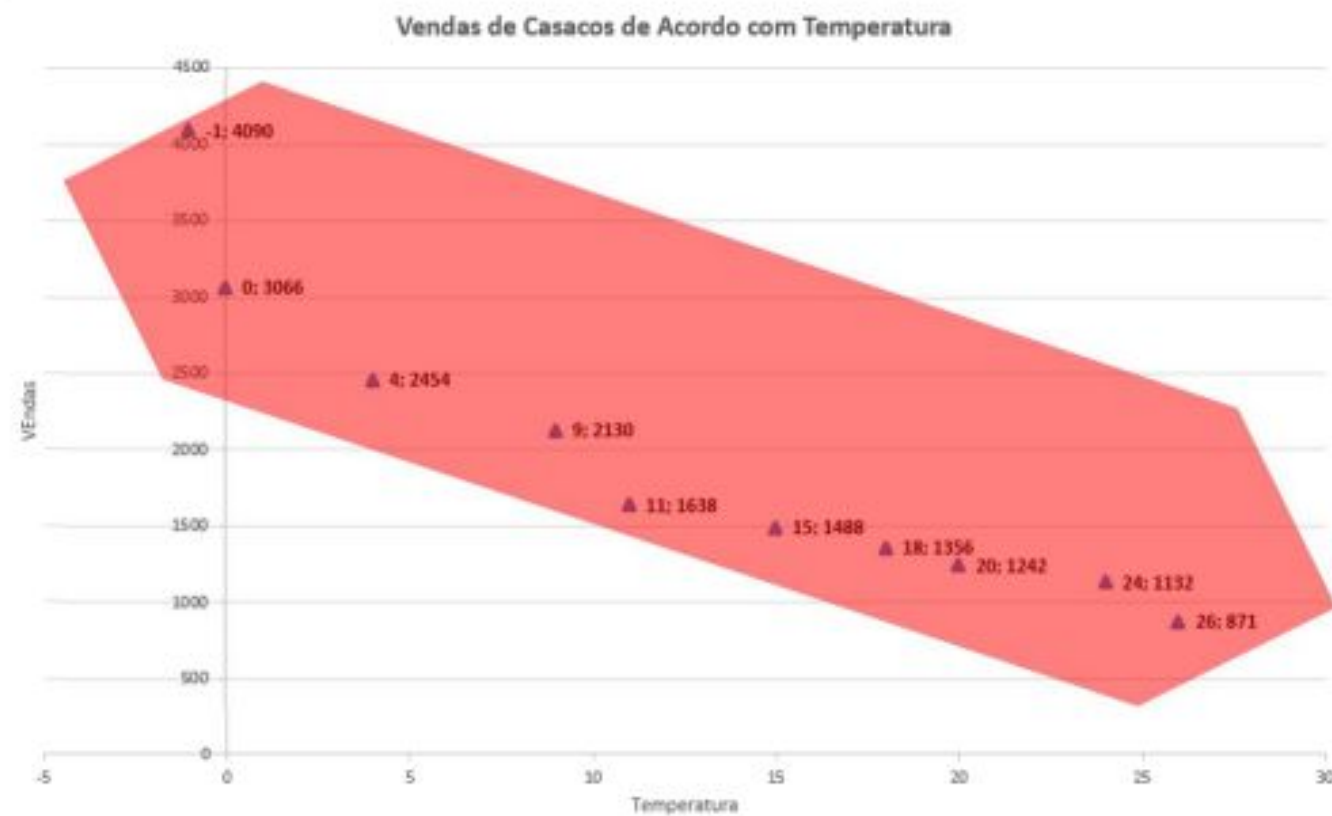


Cor: -0,93092

CORRELAÇÃO POSITIVA



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



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)



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)

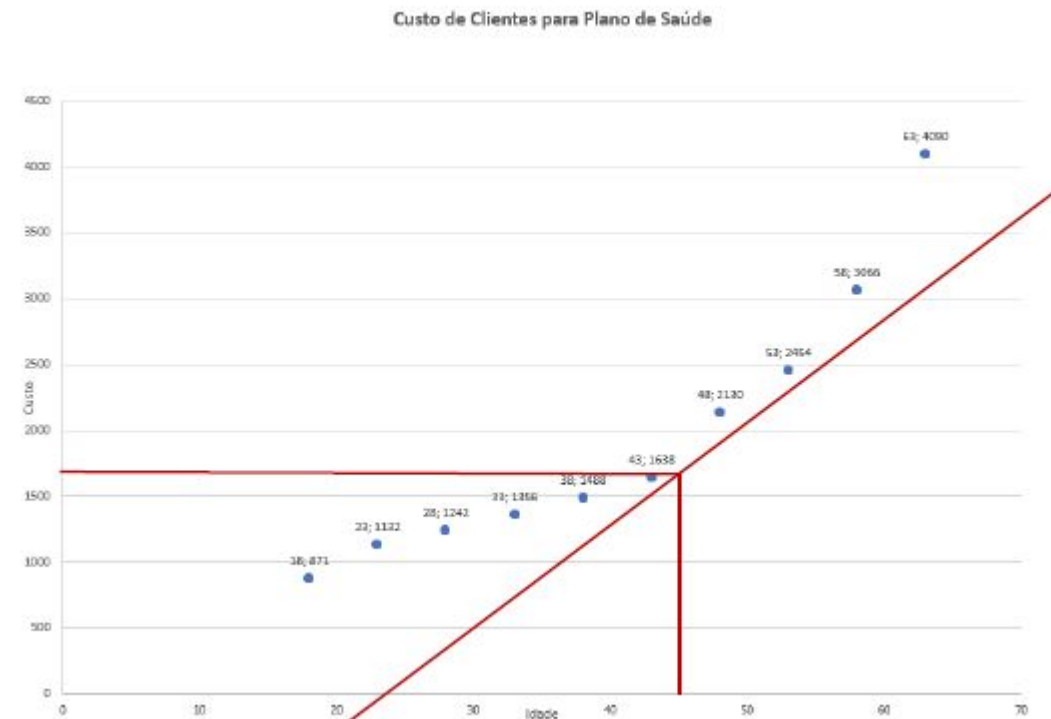
COEFICIENTE DE DETERMINAÇÃO (R^2)

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

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

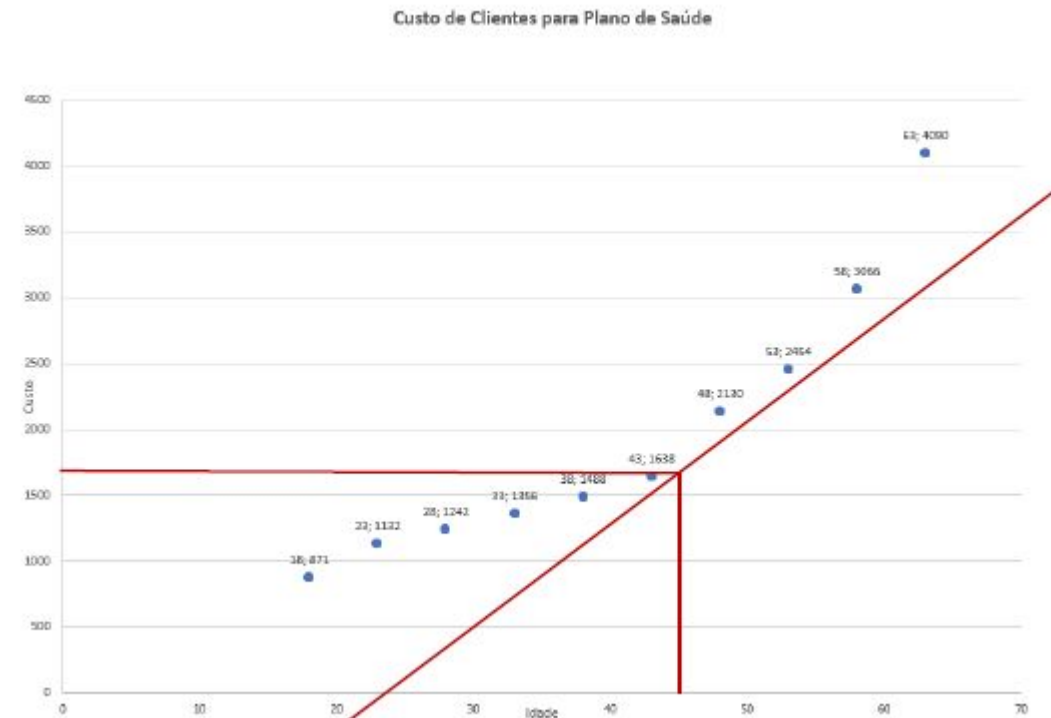
REGRESSÃO LINEAR - PREVISÃO

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



REGRESSÃO LINEAR - PREVISÃO

- Como a linha é construída?
- Ponto de Encontro da Linha no Eixo Y(interseção) : $X=0$
- Inclinação: a cada unidade que aumenta a variável Independente (x), a variável de resposta (y) sobe o valor da inclinação



REGRESSÃO LINEAR - PREVISÃO

Exemplo:

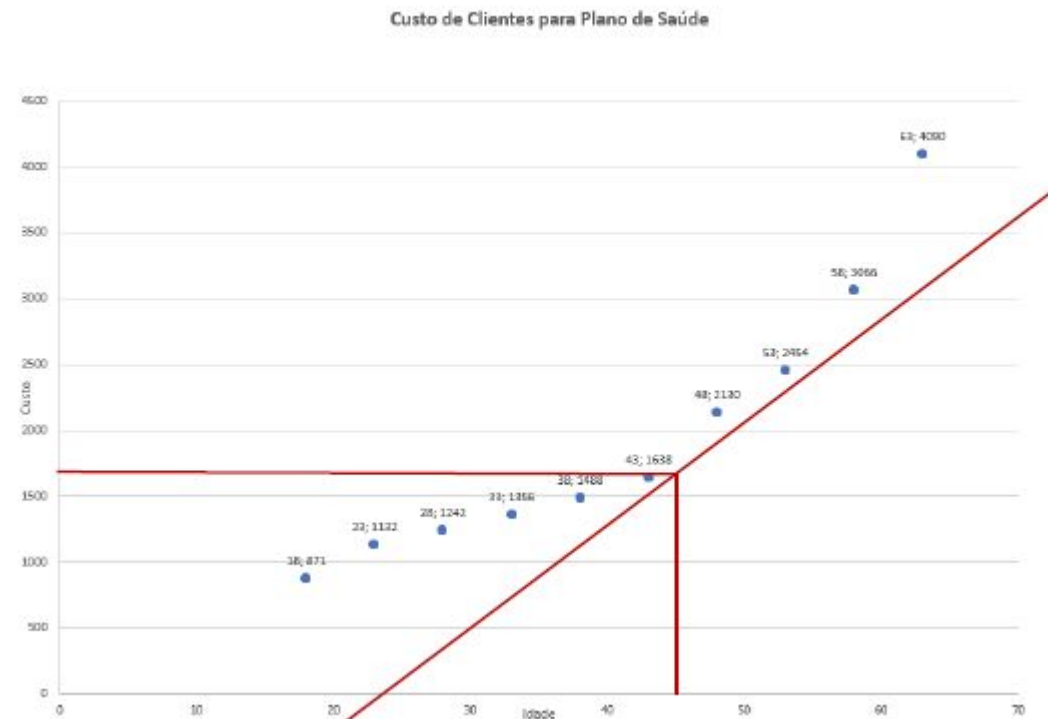
Intersecção: -558,94

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

A fórmula matemática presentes nesse modelo é a seguinte:

$$Y = A + BX$$

Variável Dependente Constantes Variável Explicativa

Sendo que na literatura estatística temos algumas outras formas de regressão que podem ser aplicadas de formas não lineares, que serão apresentadas mais à frente.

REGRESSÃO LINEAR SIMPLES

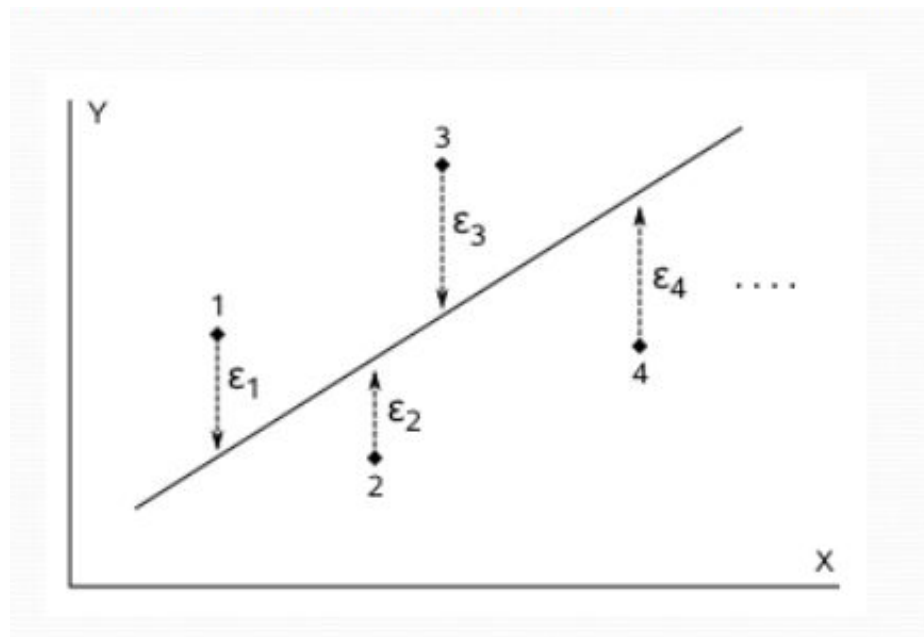
- Formalmente, a análise de regressão parte de um conjunto de observações pareadas, $(x_1, y_1), (x_2, y_2)$.. e considere que podemos escrever a relação da seguinte maneira:

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

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

ESTIMAÇÃO DE PARÂMETROS

- O objetivo é estimar valores para α e β 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

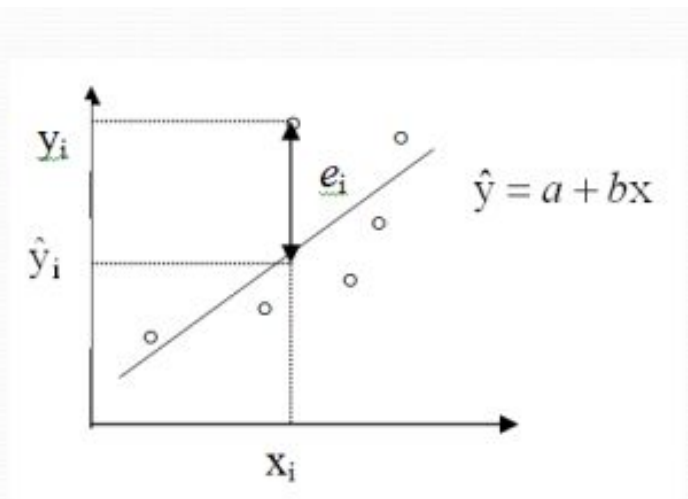
$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

RESÍDUOS

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

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

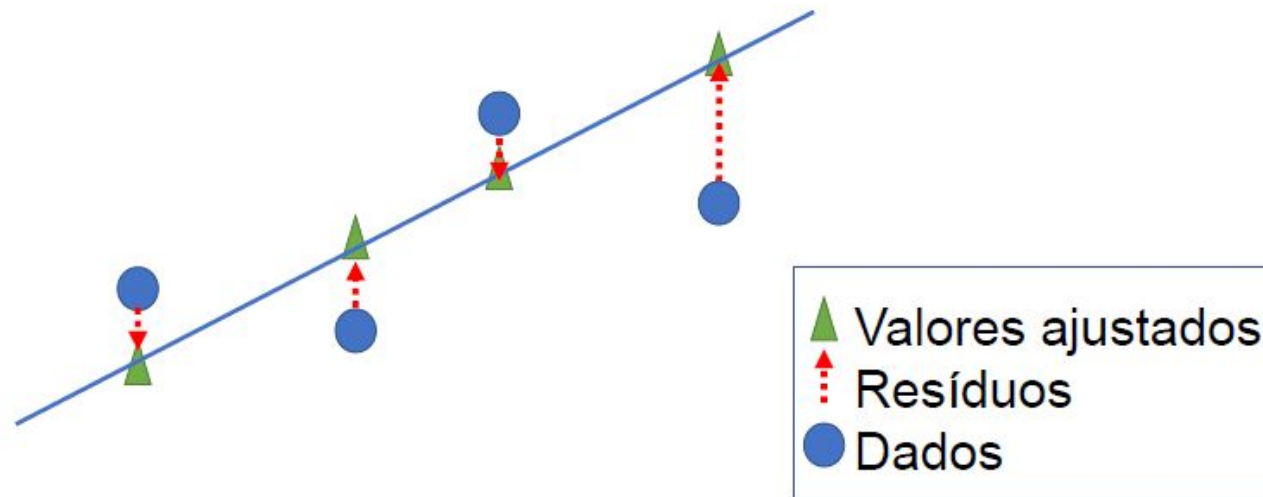
- O resíduo relativo é a i-ésima observação e pode ser chamada de erro aleatório como mostra abaixo



RESÍDUOS

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

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

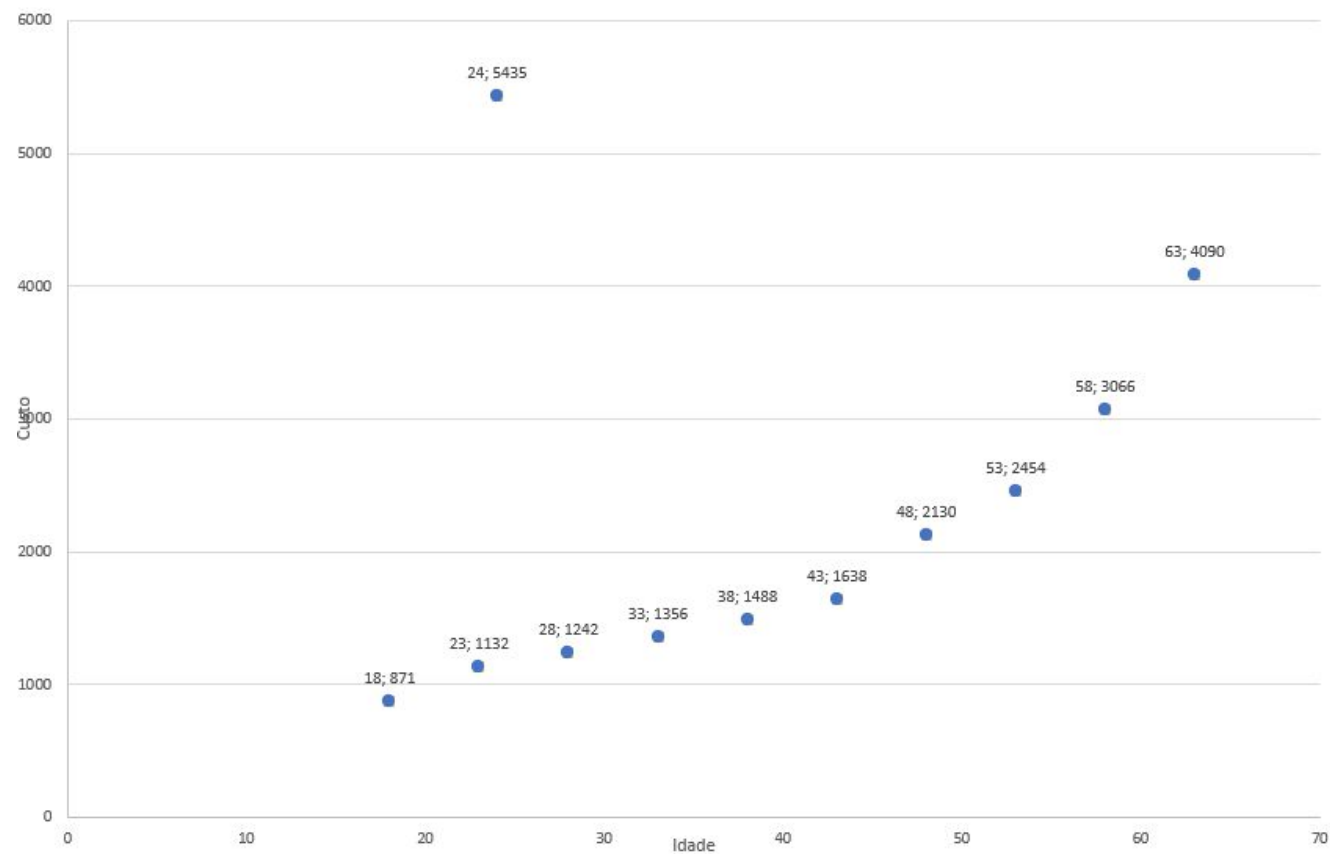


OUTLIERS

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

Nova correlação= 0,34

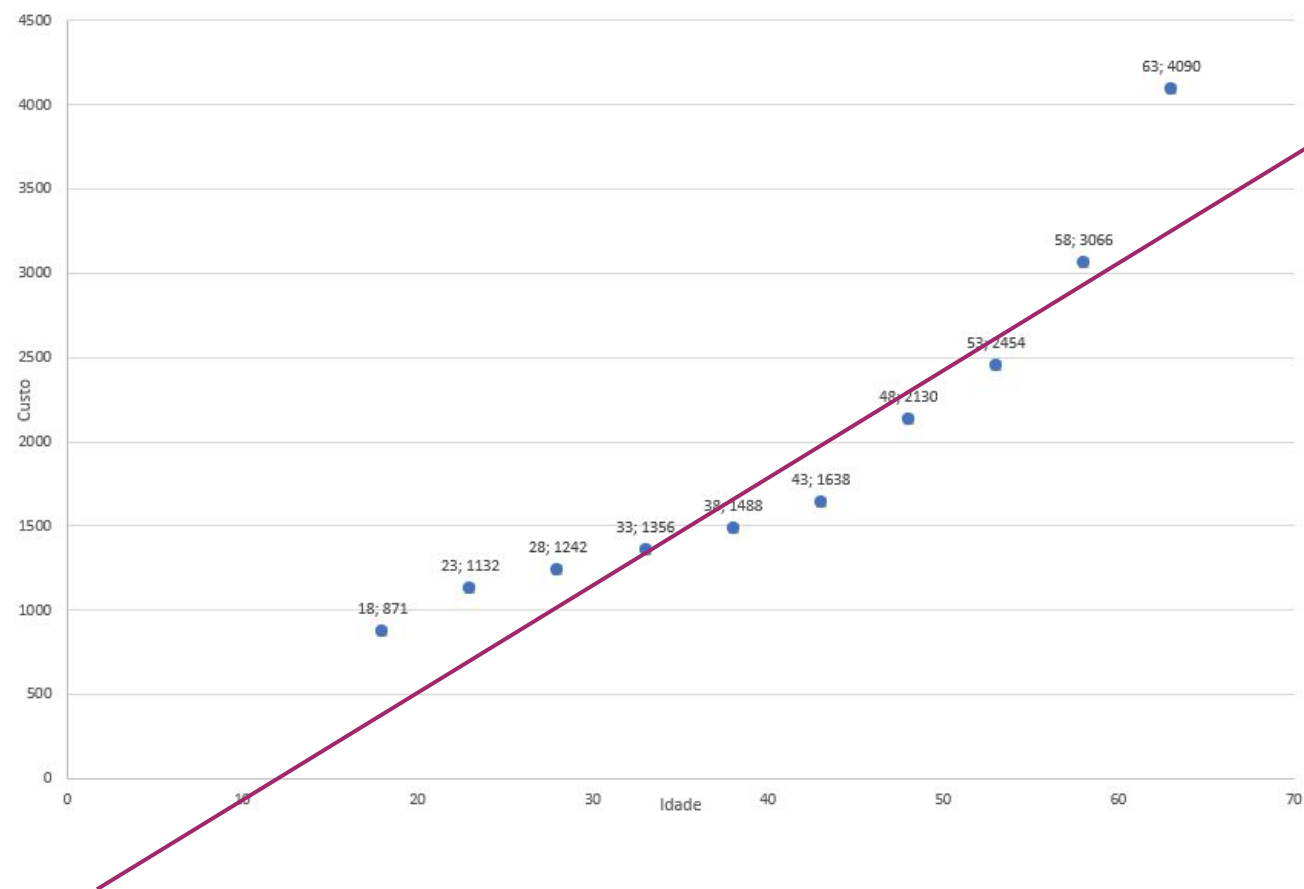
Custo de Clientes para Plano de Saúde



EXTRAPOLAÇÃO

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

Custo de Clientes para Plano de Saúde



REGRESSÃO LINEAR MÚLTIPLA

Na Simples:

Uma variável explanatória para prever uma variável dependente

$$Y \sim X$$

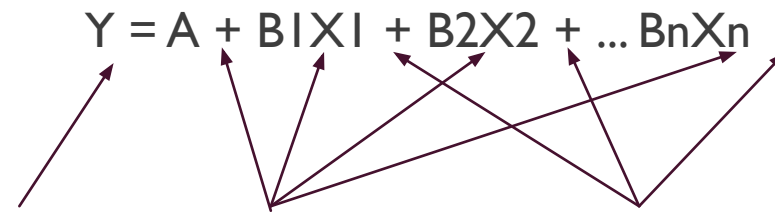
Na Múltipla

Duas ou mais variáveis explanatórias para prever uma variável dependente

$$Y \sim X_1 + X_2 + X_n$$

REGRESSÃO LINEAR MÚLTIPLA

- Em uma regressão linear múltipla temos a seguinte configuração:

$$Y = A + B_1X_1 + B_2X_2 + \dots B_nX_n$$


■ Variável Dependente Constantes Variáveis Explicativa

- Então a regressão linear múltipla nada mais é do que adicionar mais variáveis à mesma fórmula que utilizamos na regressão linear simples.
- Vejamos um exemplo.

REGRESSÃO LINEAR MÚLTIPLA

- Em python, vamos ver o exemplo dos dados de uma escola, o exemplo leva em consideração o tempo de deslocamento para a escola em relação à distância percorrida pelos alunos e os semáforos;

Estudante	Tempo para chegar à escola (minutos)	Distância percorrida (km)
Gabriela	15	8
Dalila	20	6
Gustavo	20	15
Letícia	40	20
Luiz	50	25
Leonor	25	11
Ana	10	5
Antônio	55	32
Júlia	35	28
Mariana	30	20

REGRESSÃO LINEAR MÚLTIPLA

- Então, conseguimos ver que o modelo linear simples não tão se adequou bem da forma que queríamos;
- Vamos experimentar incluir mais outra variável, bem comum no contexto de trajetos rodoviários, a quantidade de semáforos.

REGRESSÃO LINEAR MÚLTIPLA

- Exemplo: Tempo de percurso x distância percorrida e quantidade de semáforos.

Estudante	Tempo para chegar à escola (minutos)	Distância percorrida (km)	Semáforos no caminho
Gabriela	15	8	0
Dalila	20	6	1
Gustavo	20	15	0
Letícia	40	20	1
Luiz	50	25	2
Leonor	25	11	1
Ana	10	5	0
Antônio	55	32	3
Júlia	35	28	1
Mariana	30	20	1

REGRESSÃO LINEAR MÚLTIPLA

- Nossa, nosso resultado melhorou, será que devemos parar por aqui? Já é suficiente? Bom tudo determinará o quanto e onde você precisa chegar;
- Claro, pode acontecer de uma variável que for incluída no modelo não fazer parte daquele contexto e realmente ser retirada dos experimentos;
- Mas parece muito claro que, no nosso contexto, a variável quantidade de semáforos no caminho pode ser sim determinante no nosso modelo.

REGRESSÃO LINEAR MÚLTIPLA

- Porém até agora em nossos estudos só trabalhamos com variáveis quantitativas, ou seja, que possuem um valor numérico;
- Quando estivermos trabalhando com variáveis qualitativas, não podemos simplesmente criar uma escala qualquer pois isso geraria um erro grave, que a literatura denomina como “ponderação arbitrária”;
- Para isso, existe o artifício das variáveis dummy, ou binárias, que tentam abstrair qualidades em forma de 0 ou 1 para expressar determinada condição qualitativa.

REGRESSÃO LINEAR MÚLTIPLA

- Imaginemos então que vamos adicionar mais outra variável ao nosso contexto, agora uma variável qualitativa;
- Vamos incluir o período do dia no qual o trajeto foi feito, e considerando que a aula só acontece em dois períodos a variável qualitativa seguirá a seguinte lógica binária:
 - 1 – Manhã;
 - 0 – Tarde.
- Portanto, vamos ver como essa informação fica dentro da tabela.

REGRESSÃO LINEAR MÚLTIPLA

- Exemplo: Tempo de percurso x distância percorrida e quantidade de semáforos.

Estudante	Tempo para chegar à escola (minutos)	Distância percorrida (km)	Semáforos no caminho	Período do dia
Gabriela	15	8	0	1
Dalila	20	6	1	1
Gustavo	20	15	0	1
Letícia	40	20	1	0
Luiz	50	25	2	0
Leonor	25	11	1	1
Ana	10	5	0	1
Antônio	55	32	3	0
Júlia	35	28	1	1
Mariana	30	20	1	1

REGRESSÃO LINEAR MÚLTIPLA

Analisar Cada x Com y

- Analisar cada variável independente com y individualmente
- Gerar gráficos de dispersão individuais
- Buscar redundâncias (mesmos efeitos de x sobre y)

REGRESSÃO LINEAR MÚLTIPLA

- Coeficiente de Determinação (R^2)

Lembrando que R^2 é o percentual de variação da variável de resposta que é explicada pelo modelo

Quando se colocam mais variáveis no modelo, **a tendência é que R^2 aumente**, mesmo que a adição da variável não aumente a precisão do modelo

Para isso, utiliza-se R^2 ajustado, que ajusta a variação do modelo de acordo com o número de variáveis independentes que é incluída no modelo

- R^2 ajustado vai ser sempre menor que R^2

QUANDO USAR CORRELAÇÃO LINEAR

CORRELAÇÃO DE MODERADA A FORTE

$R^2 > 0.7$

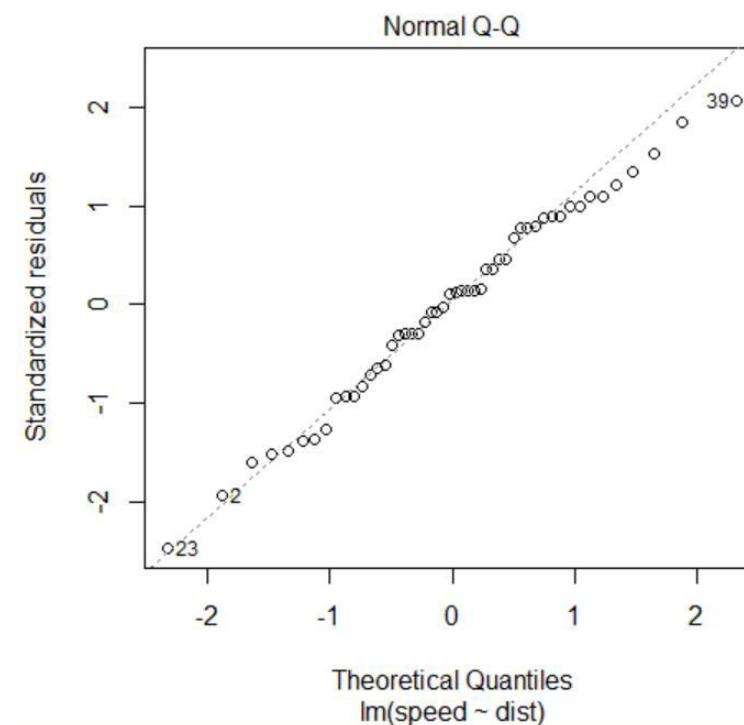
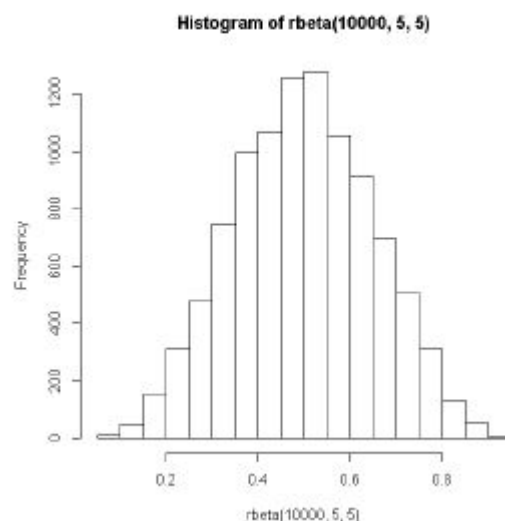
QUANDO USAR: RESIDUAIS PADRONIZADOS

Residuais próximos de uma distribuição normal:

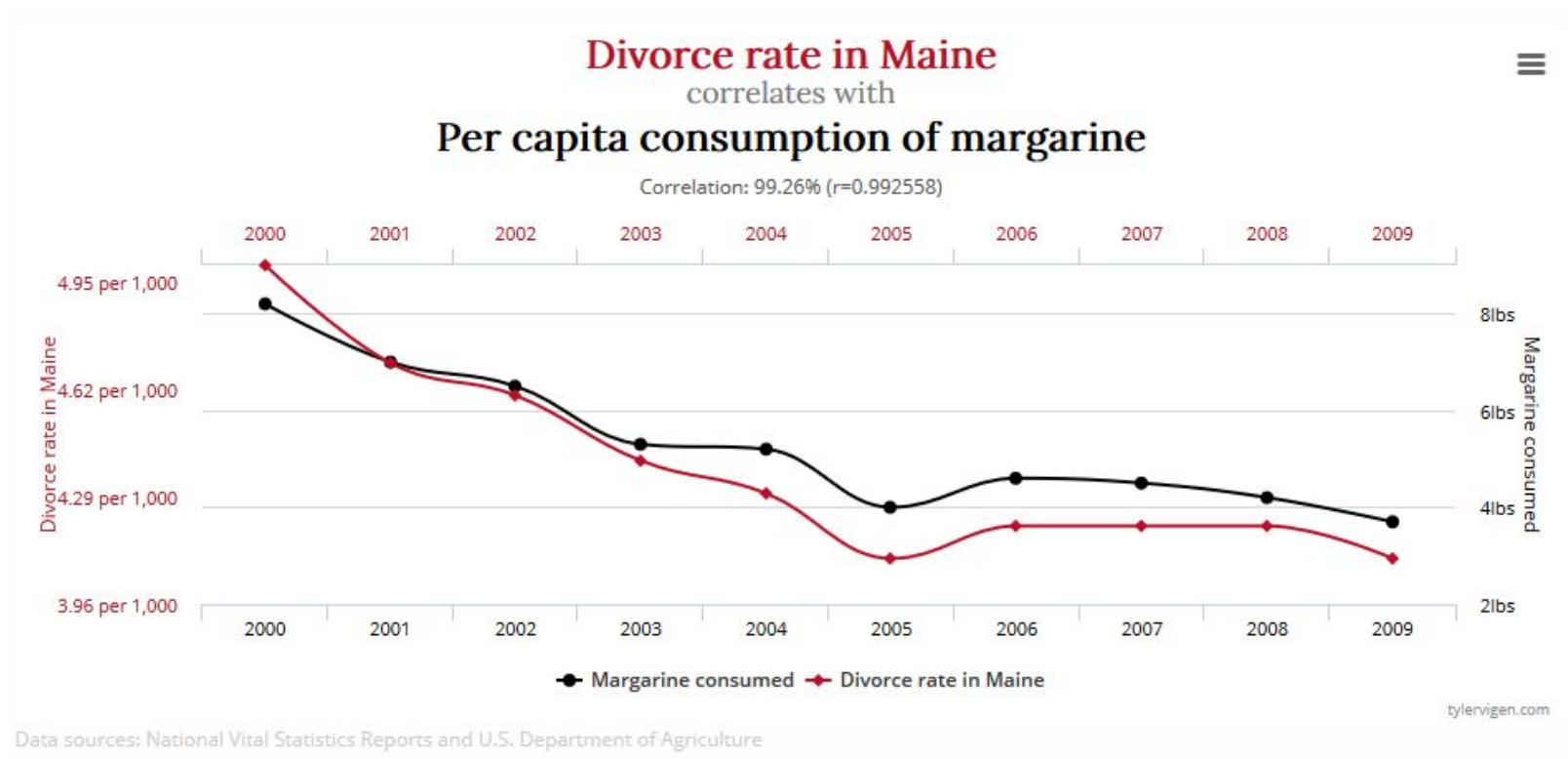
Histograma

Diagrama de normalidade

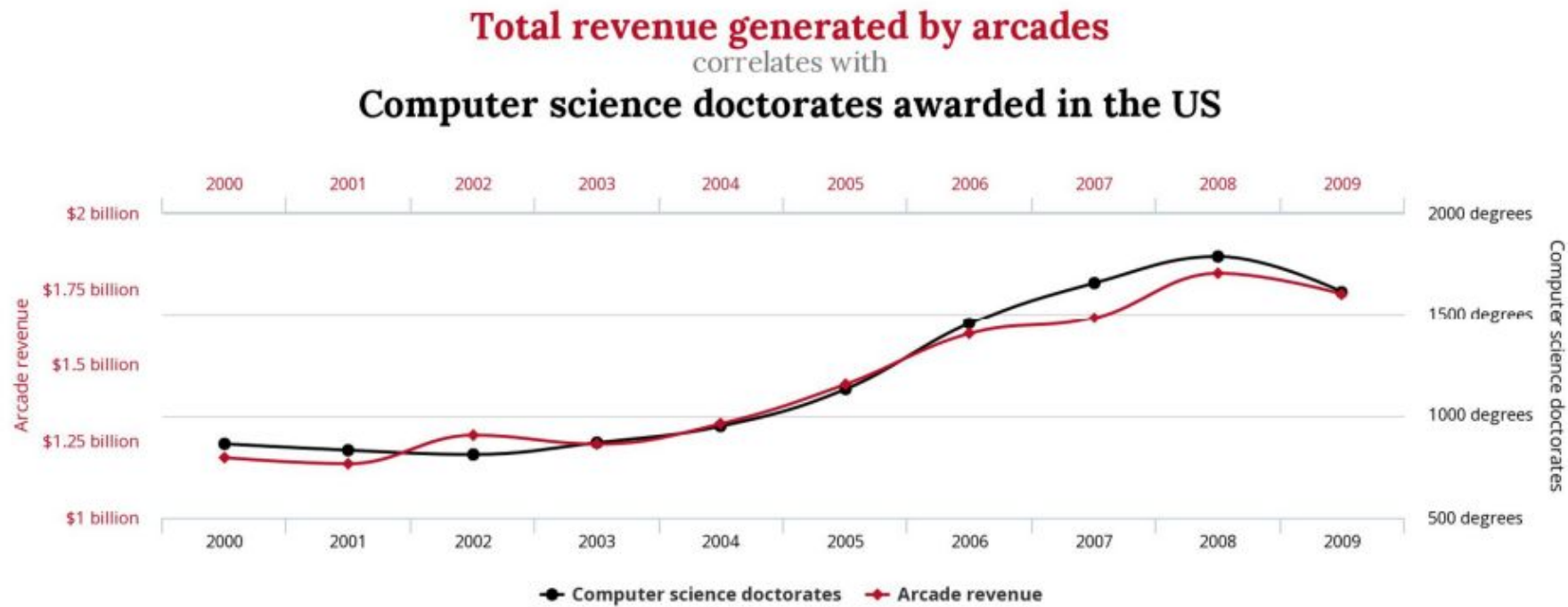
Teste de Shapiro-Wilk



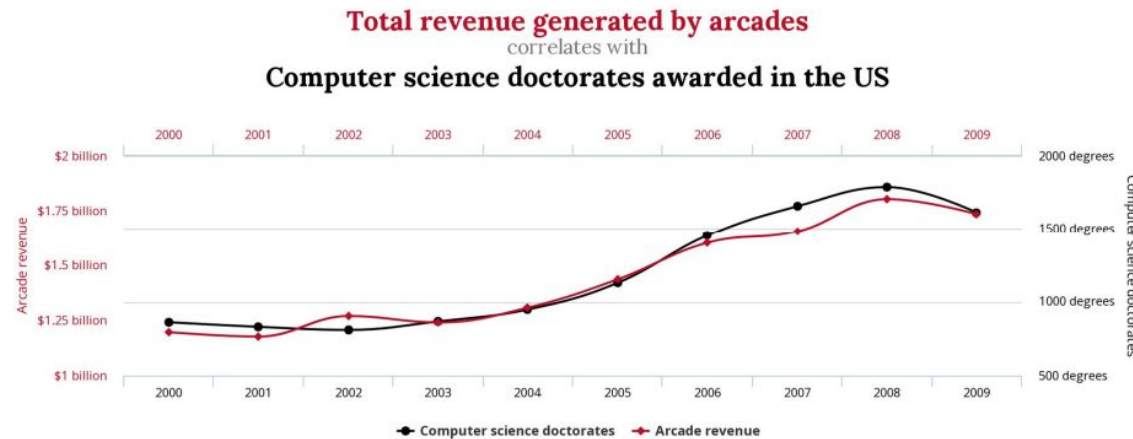
CORRELAÇÃO NÃO É CAUSA



CORRELAÇÃO NÃO É CAUSA



CORRELAÇÃO NÃO É CAUSA



- Podemos afirmar ?
 - As vendas de Arcade geram receita para que estudantes de TI façam doutorado
 - As pessoas de TI que fazem doutorado, investem em Arcade para obter receita

CORRELAÇÃO NÃO É CAUSA

- Os exemplos anteriormente mostrados são do site Spurious Correlations, que pode ser acessados pelo link: <http://tylervigen.com/spurious-correlations>
- Esses casos que são todos reais, mostra de forma absurda, mas extremamente visual que o fato de dados serem correlacionados não necessariamente indica causalidade;
- E como provamos a causalidade então? Bom, como engenheiros em formação vocês já devem imaginar a resposta: Experimentando;