Preditiva.ai

Live #10

Os principais erros em Análise de Dados e como evitá-los

Parte 1

21/02/2023

Antes de mais nada, já se inscreveu em nosso canal para ter acesso aos materiais e avisos das lives?

https://t.me/preditiva

Erros acontecem...

A definição formal

Preditiva.ai

Risco operacional

Probabilidade de ocorrência de perdas resultantes de eventos externos ou de falha, deficiência ou inadequação de processos internos, pessoas ou sistemas.

Erros acontecem...

A definição que eu gosto é...

Preditiva.ai

Risco operacional

=

Probabilidade de você fazer uma CAGADA!

Jornada dos dados

Existem muitas chances de erros na jornada dos dados

Preditiva.ai

Dados brutos

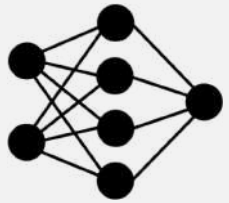
Processo de Transformação dos Dados

Dados precisam ser organizados, limpos, padronizados para serem armazenados no repositório

Data Lake (repositório)

Decisões são tomadas
Desenvolvimento de análises exploratórias e Dashboards interativos

Analytics e Dashboards



Machine Learning e AI

Decisões automáticas são tomadas
Desenvolvimento de modelos estatísticos e de Machine Learning/AI

Produção/Deploy

Produtização da decisão
Construção de códigos que asseguram uma eficiente implantação dos produtos de dados anteriores

Valor Tangível que os dados podem entregar

Metodologia de Projetos de Dados e Data Storytelling

Jornada dos dados

Existem muitas chances de erros na jornada dos dados

Preditiva.ai

Dados brutos

Processo de Transformação dos Dados

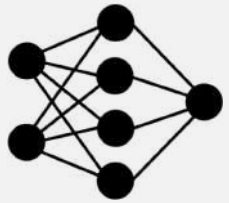
Dados precisam ser organizados, limpos, padronizados para serem armazenados no repositório

Data Lake (repositório)

Decisões são tomadas
Desenvolvimento de análises exploratórias e Dashboards interativos

Analytics e Dashboards

Live - Parte 1



Machine Learning e AI

Decisões automáticas são tomadas
Desenvolvimento de modelos estatísticos e de Machine Learning/AI

Produção/Deploy

Produtização da decisão
Construção de códigos que asseguram uma eficiente implantação dos produtos de dados anteriores

Valor Tangível que os dados podem entregar

Live - Parte 2

Metodologia de Projetos de Dados e Data Storytelling

O mapa da CAGADA em Dados!

Parte 1

Preditiva.ai

Medidas e Gráficos

- Medidas estatísticas
- Unidades da base
- Medidas vs Gráficos
- Correlações
- Causalidade

Projeções

- Frequências imprecisas
- Probabilidade e Possibilidade

Amostragem

- Viés de seleção
- Erros de Coleta
- Tamanho de Amostra

O mapa da CAGADA em Dados!

Parte 1

Preditiva.ai

Medidas e Gráficos

- **Medidas estatísticas**
- **Unidades da base**
- **Medidas vs Gráficos**
- **Correlações**
- **Causalidade**

Projeções

- Frequências imprecisas
- Probabilidade e Possibilidade

Amostragem

- Viés de seleção
- Erros de Coleta
- Tamanho de Amostra

Medidas e Gráficos

Exemplo 1

Preditiva.ai

O que você faria na seguinte situação?

Você é convidado para trabalhar em uma startup com 15 funcionários e, segundo o RH o salário **médio** dos funcionários é R$4.200,00.

Você atualmente ganha R$1.000,00.

Funcionário	Salário
1	1.000,00
2	1.000,00
3	1.000,00
4	1.000,00
5	1.000,00
6	1.000,00
7	1.000,00
8	1.000,00
9	1.000,00
10	1.000,00
11	1.000,00
12	1.000,00
13	1.000,00
14	10.000,00
15	40.000,00
Média	**4.200,00**

Medidas e Gráficos

Medidas Resumo: Medidas de Posição - Média

Preditiva.ai

ID	Salário
1	5.130,00
2	4.193,00
3	3.468,00
4	3.068,00
5	2.670,00
6	2.693,00
7	9.526,00
8	3.068,00
9	5.237,00
10	9.980,00
11	2.426,00
12	2.911,00

A **Média** é uma **medida de tendência central**. Isto é, ela tenta nos dar uma noção de onde o valor central dos dados deve aparecer. Veja:

$$\text{Média} = \frac{54.370,00}{12} = 4.530,83$$

=MÉDIA()

Porém, muitas vezes **ela não é a medida resumo mais indicada**. Veremos isso mais adiante.

Medidas e Gráficos

Medidas Resumo: Medidas de Posição - Mediana

ID	Salário
11	2.426,00
5	2.670,00
6	2.693,00
12	2.911,00
4	3.068,00
8	3.068,00
3	3.468,00
2	4.193,00
1	5.130,00
9	5.237,00
7	9.526,00
10	9.980,00

Mediana

3.268,00

A **Mediana** , assim como a média, também é uma medida de tendência central. Porém, no seu caso, ela realmente representa o centro do seu conjunto de dados.

Ou seja, 50% dos valores são inferiores à **Mediana** e 50% dos valores são superiores à **Mediana**.

Obs: Quando o número de observações é par (no exemplo, temos 12 salários), a **Mediana** é a média entre os 2 valores centrais. Se fosse ímpar, seria o próprio valor central.

=MED()

Medidas e Gráficos

Medidas Resumo: Medidas de Posição

Preditiva.ai

O que você faria na seguinte situação?

Você está doente e só existe um remédio que pode te ajudar. Segundo a bula, o tempo de sobrevida **mediano** é de 8 semanas. Todos que tomam o remédio tem severos efeitos colaterais. Vale a pena tomar o remédio?

Tempo de Sobrevida (em semanas)	% Relativa
2	15%
4	15%
8	20%
16	5%
32	5%
64	5%
128	5%
256	10%
512	20%

Medidas e Gráficos

Exemplo 1 - Resumo

Preditiva.ai

O erro:

Usar apenas **uma** medida para resumir os dados.

Como evitá-lo?

Quanto mais medidas resumo usar, melhor conseguirá entender os dados analisados. Exemplos de medidas:

- Média
- Mediana
- Quartis
- Mínimo e Máximo
- Desvio Padrão

Medidas e Gráficos

Exemplo 2

Preditiva.ai

Você pede para alguém um relatório de vendas anual. O analista entrega isso...

Conclusões mais comuns:

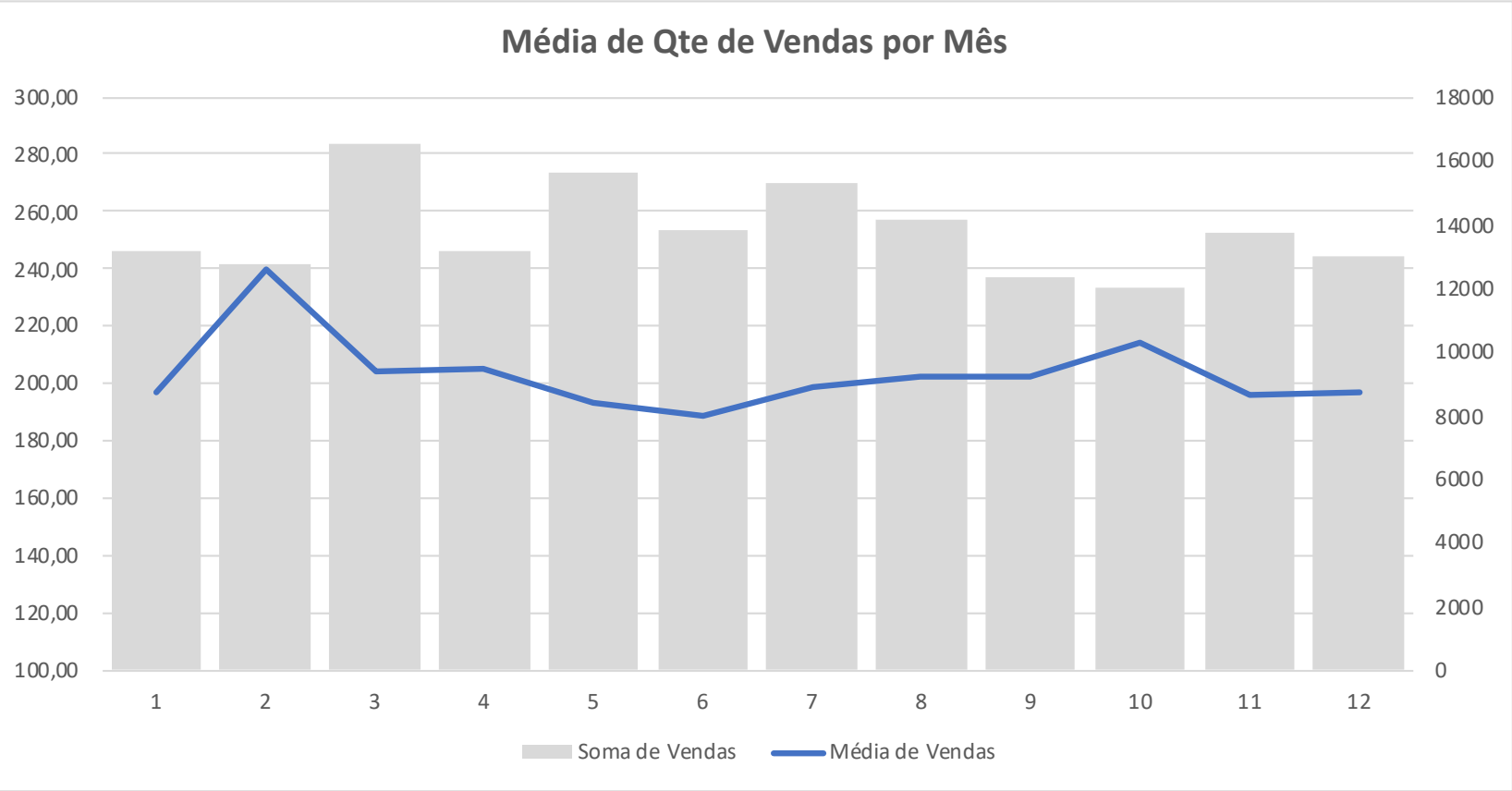
1. Fevereiro foi um mês atípico de vendas.
2. De março a setembro ficamos estáveis em cerca de 200 vendas por mês.

Medidas e Gráficos

Exemplo 2

Preditiva.ai

Para tentar entender melhor, você pede a soma de vendas por mês. O analista te entrega isso...



Sua conclusão mudou de alguma forma?

Será que conseguimos realmente entender nossas vendas com esse resumo?

E se tivéssemos acesso aos dados brutos para um *Double Check*?

Medidas e Gráficos

Exemplo 2 - Resumo

Preditiva.ai

O erro:

Assumir a unidade da base utilizada sem perguntar para o analista gerador da análise.

Como evitá-lo?

Deixar claro qual unidade da base deve ser trabalhada (tanto se você for o cliente do trabalho quanto você for o criador da análise).

Pergunte também:

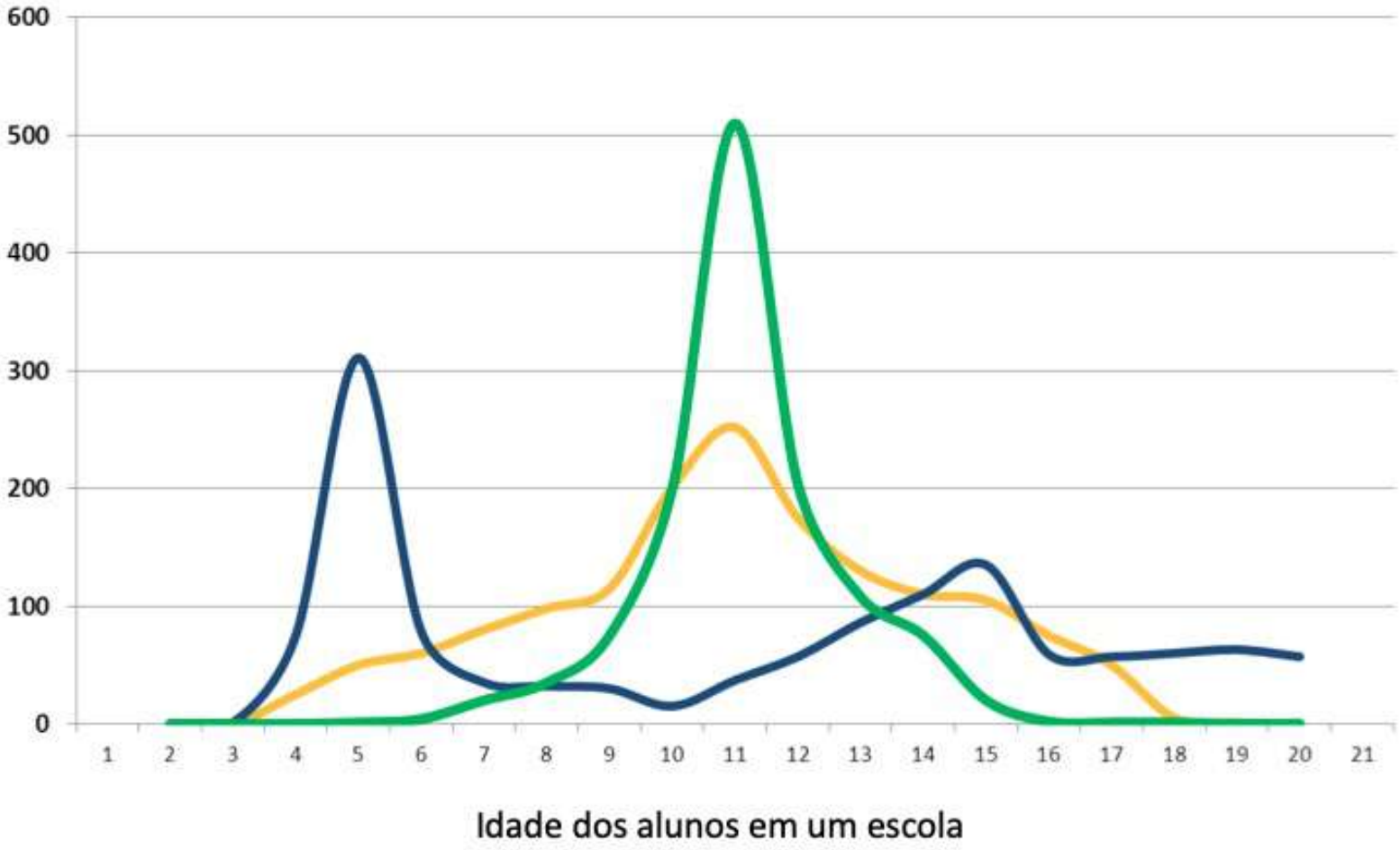
- Filtro de tempo usado (Diário, Semanal, Mensal ?)
- Se existem outros tipos de filtros ocultos (ex: produto, segmentos, regiões etc)

Medidas e Gráficos

Exemplo 3

Preditiva.ai

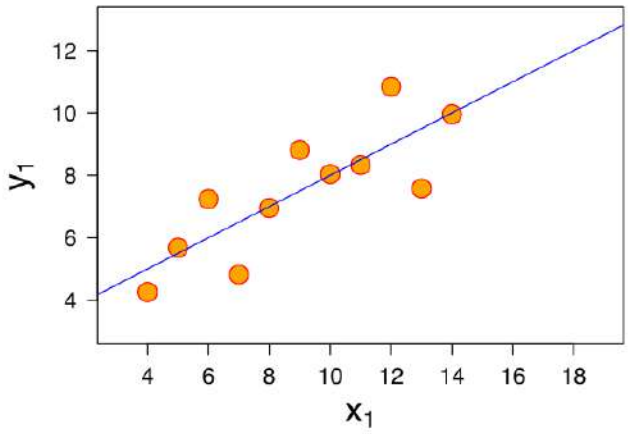
Qual linha tem a maior média?

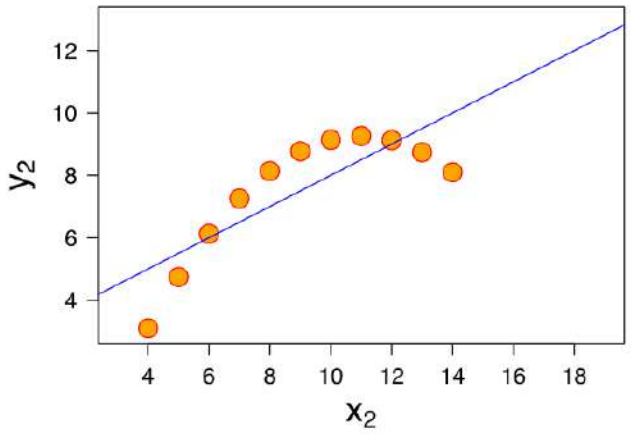


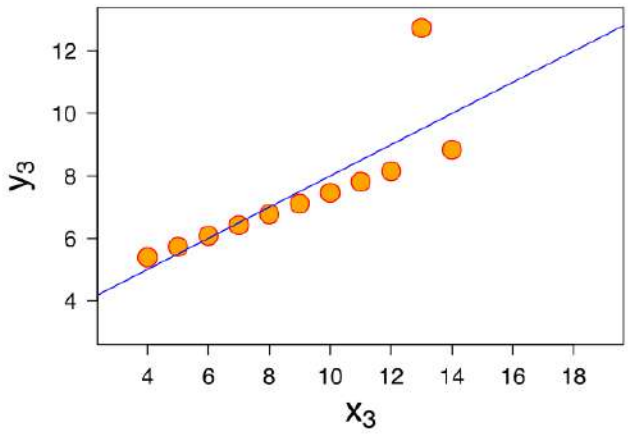
Medidas e Gráficos

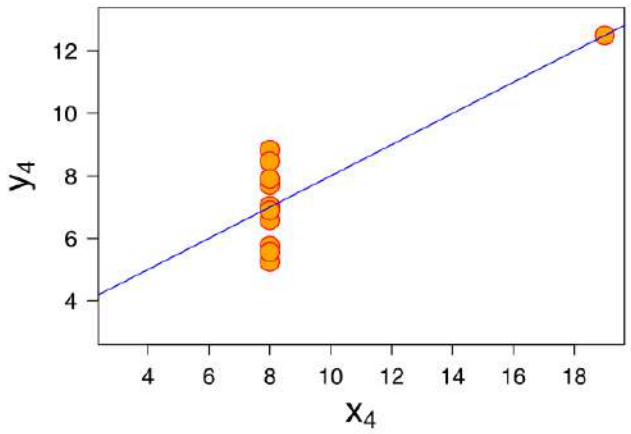
Exemplo 4 - O que esses quatro conjuntos têm em comum?

Preditiva.ai









Propriedade	Valor
Média em x	9
Variância em x	11
Média em y	7.50
Variância em y	4.125
Correlação entre x e y	0.816
Regressão linear	$y = 3.00 + 0.500x$
R^2	0.67

Quarteto de Ascombe

Medidas e Gráficos

Exemplo 3 e 4 - Resumo

Preditiva.ai

O erro:

As vezes nem muitas medidas resumo são suficientes para entender seus dados.

Como evitá-lo?

Sempre que possível plotar os dados em um gráfico analítico e comparar com as medidas resumo.

Gráficos mais indicados:

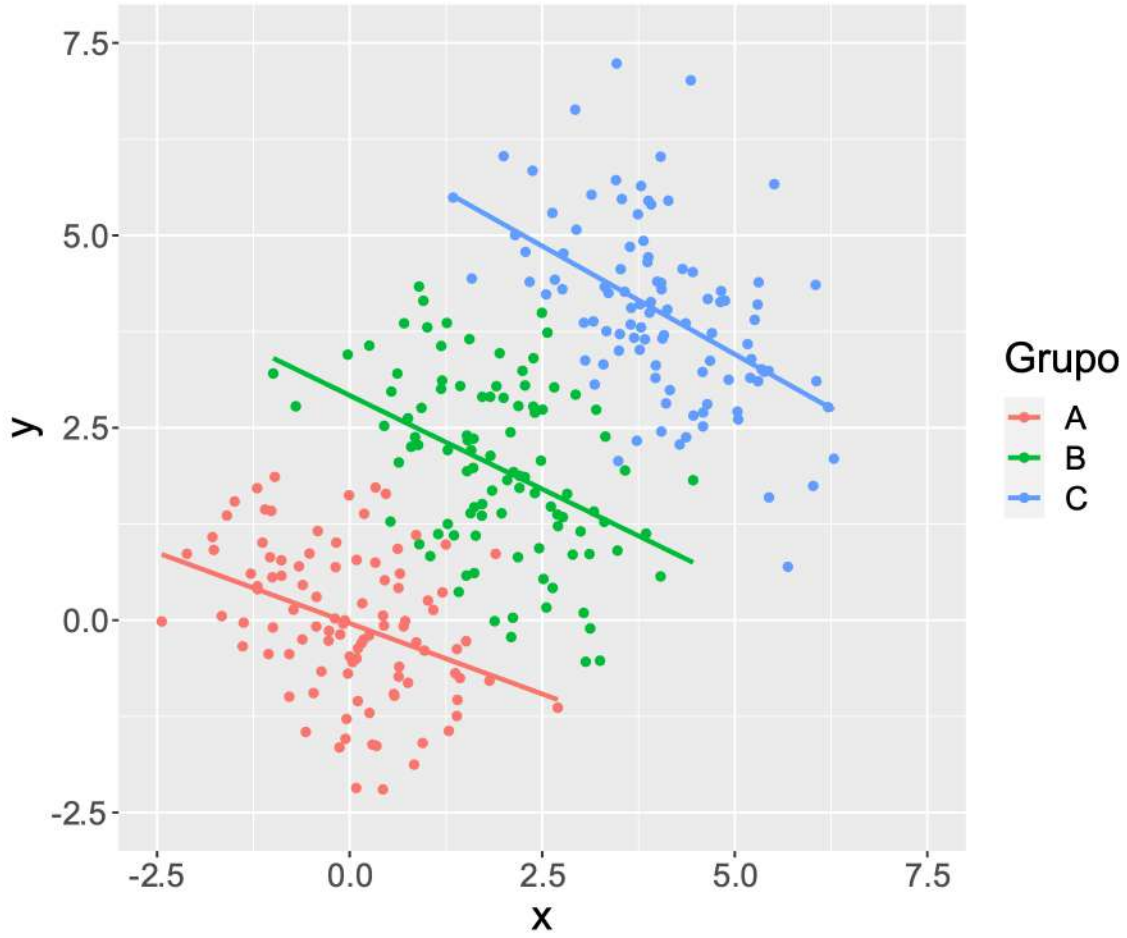
- Histogramas
- Boxplots
- Gráficos de Dispersão (Scatter Plots)

Medidas e Gráficos

Exemplo 5

Preditiva.ai

Paradoxo de Simpson



Medidas e Gráficos

Exemplo 5 - Resumo

Preditiva.ai

O erro:

A correlação entre seus dados pode mudar ao agrupá-los de formas diferentes.

Como evitá-lo?

Sempre plote o gráfico para cada grupo mais importante de sua base de dados.

Agrupamentos possíveis:

- Por tempo
- Por produto
- Por região
- Por dados demográficos

Medidas e Gráficos

Exemplo 6

Preditiva.ai

Você pede para um analista um relatório de vendas de um determinado produto. A ideia é saber se o gênero explica a compra desse produto. O analista te entrega isso:

Gênero	Unidades Vendidas	**Frequência de Compra (%)**
Masculino	400	20%
Feminino	1600	80%
	2000	

Conclusões mais comuns:

Produto claramente voltado para o público feminino.

Medidas e Gráficos

Exemplo 6

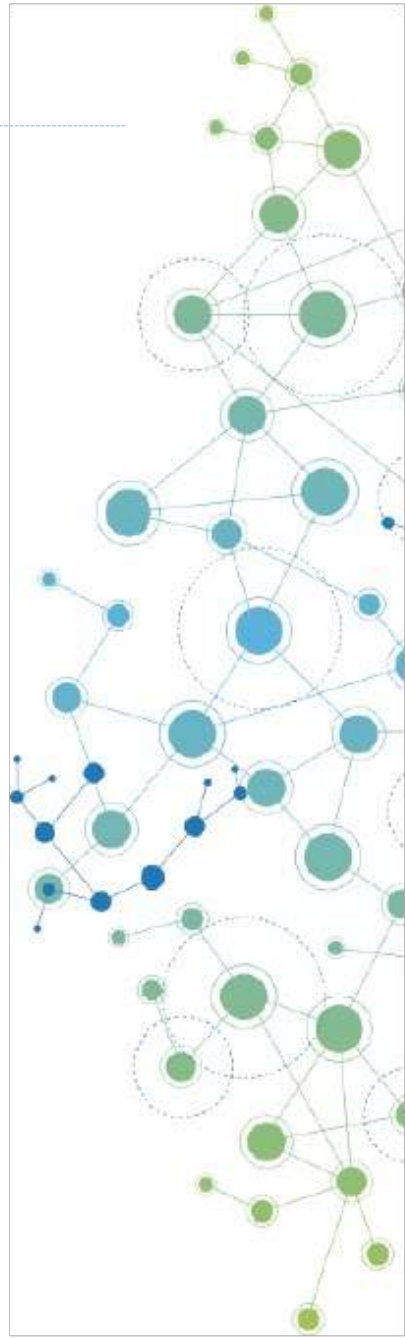
Preditiva.ai

Essa análise simples não considera o fator **contrafactual**, ou seja, não considera as compras **não** realizadas por cada um dos gêneros. Teria a mesma proporção? Vejamos:

Gênero	Unidades enviadas para o carrinho de compras	**Frequência de Compra (%)**	**Frequência de Não Compra (%)**	**Diferença**
Masculino	3600	20%	19%	1%
Feminino	7200	80%	81%	-1%
	6000			

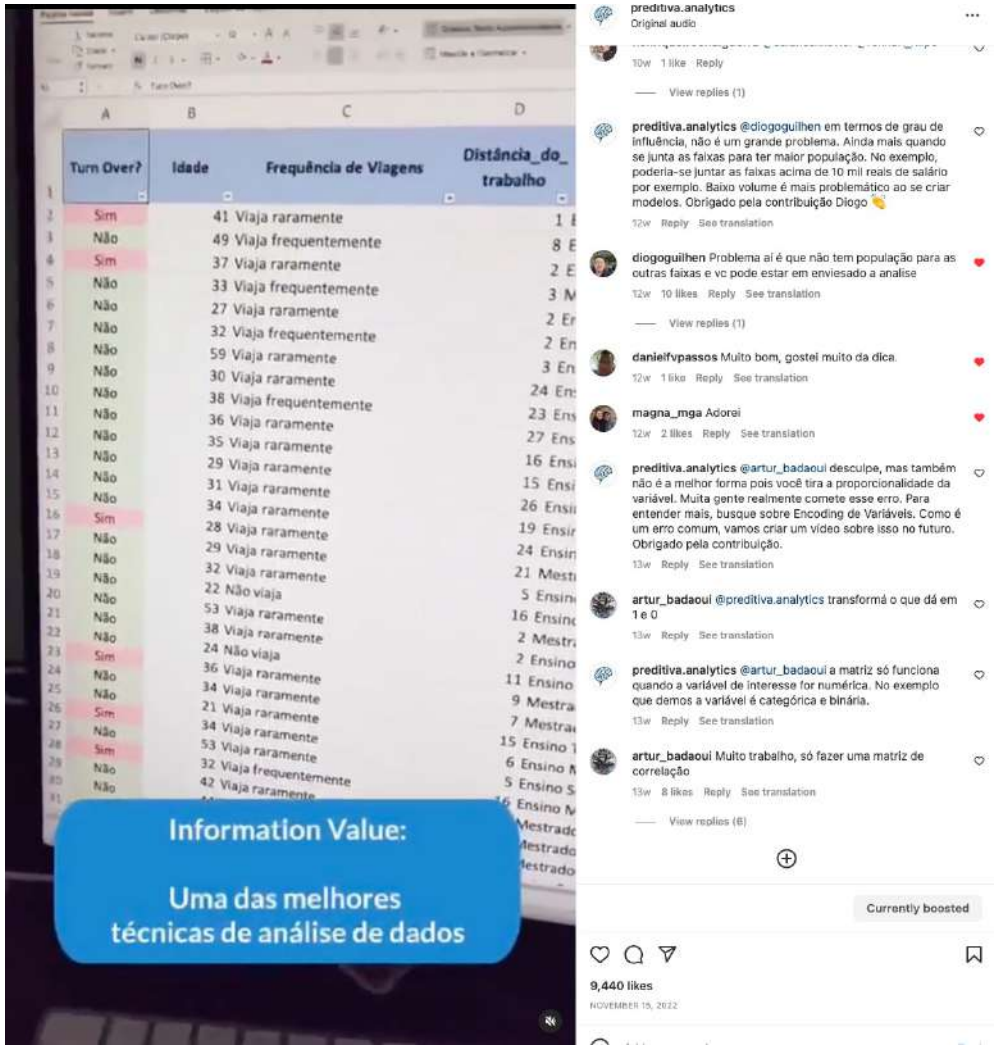
E agora? A conclusão muda?

Demonstração desse erro



Medidas e Gráficos

Exemplo 7 – Correlação de Pearson não é a única forma de medir



preditiva.analytics
Original audio

preditiva.analytics @diogoguilhen em termos de grau de influência, não é um grande problema. Ainda mais quando se junta as faixas para ter maior população. No exemplo, poderia-se juntar as faixas acima de 10 mil reais de salário por exemplo. Baixo volume é mais problemático ao se criar modelos. Obrigado pela contribuição Diogo

diogoguilhen Problema aí é que não tem população para as outras faixas e vc pode estar em enviesado a analise

danielfvpassos Muito bom, gostei muito da dica.

magna_mga Adorei

preditiva.analytics @artur_badaoui desculpe, mas também não é a melhor forma pois você tira a proporcionalidade da variável. Muita gente realmente comete esse erro. Para entender mais, busque sobre Encoding de Variáveis. Como é um erro comum, vamos criar um vídeo sobre isso no futuro. Obrigado pela contribuição.

artur_badaoui @preditiva.analytics transformá o que dá em 1 e 0

preditiva.analytics @artur_badaoui a matriz só funciona quando a variável de interesse for numérica. No exemplo que demos a variável é categórica e binária.

artur_badaoui Muito trabalho, só fazer uma matriz de correlação

Currently boosted

9,440 likes

NOVEMBER 15, 2022

[redacted] Muito trabalho, só fazer uma matriz de correlação

13w 8 likes Reply See translation

preditiva.analytics [redacted] a matriz só funciona quando a variável de interesse for numérica. No exemplo que demos a variável é categórica e binária.

13w Reply See translation

[redacted] @preditiva.analytics transformá o que dá em 1 e 0

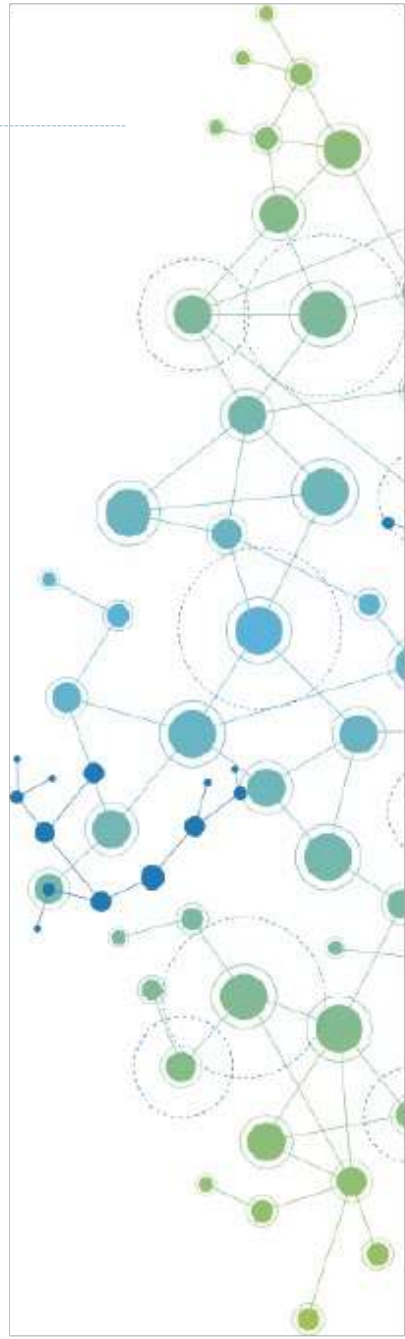
13w Reply See translation

[redacted] Faz um heatmap da Matriz de Correlação das variaveis po

6w Reply See translation

[redacted] Muito trabalho, só fazer uma matriz de correlação

Demonstração desse erro



Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

É fundamental dominarmos a **diferença entre esses 2 conceitos** para não cairmos em algumas **armadilhas** de Analytics.

Vejamos a definição destes 2 termos:

- **Correlação**: relação de **dependência** ou **associação** entre duas variáveis.
- **Causalidade**: relação entre um evento A e um evento B, sendo que o evento B é **consequência** do evento A.

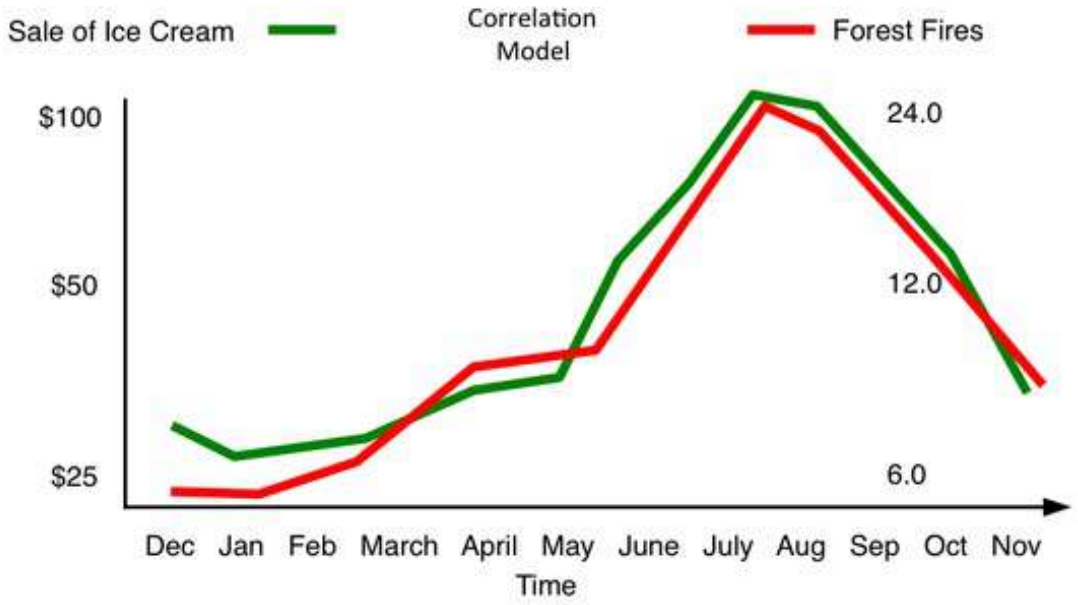
Ou seja, **Correlação** está relacionada com a **dependência ou associação** e a **Causalidade** relacionada a **consequência**.

Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Vamos avaliar a **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas**:



Fonte: https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect

Você acha que a **venda de sorvetes** pode **causar incêndios nas florestas**?

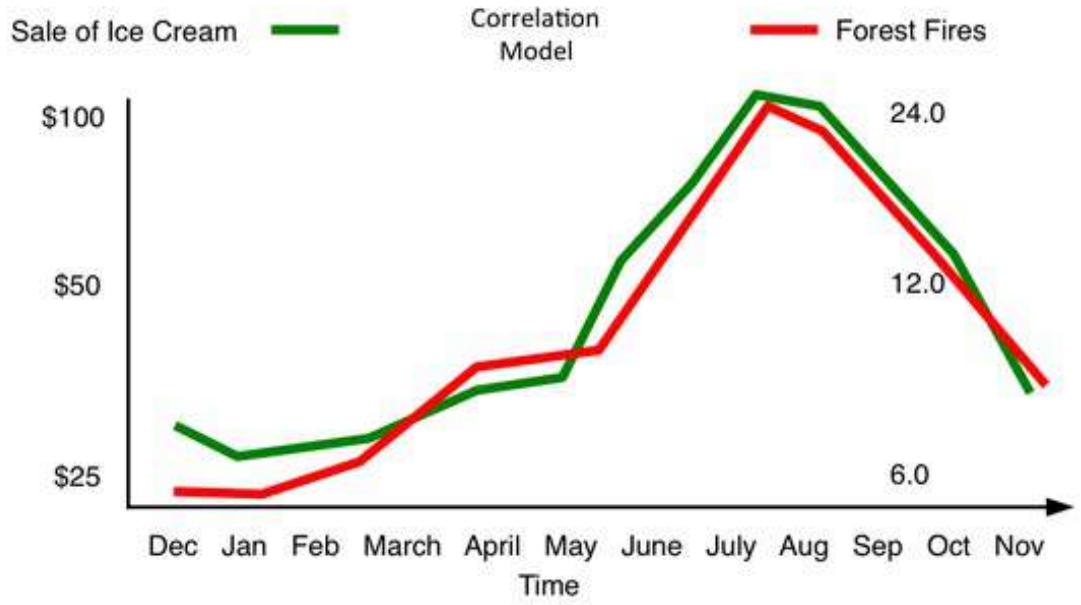
Neste caso há uma **3ª variável não avaliada** e que faz mais sentido ser a **causadora** do aumento no **consumo de sorvete** e dos **incêndios nas florestas**: **o clima quente!**

Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Vamos avaliar a **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas**:



Fonte: https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect

A **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas** é conhecida como **Correlação Espúria**.

As **Correlações Espúrias** podem ser uma armadilha para **falsas conclusões**.

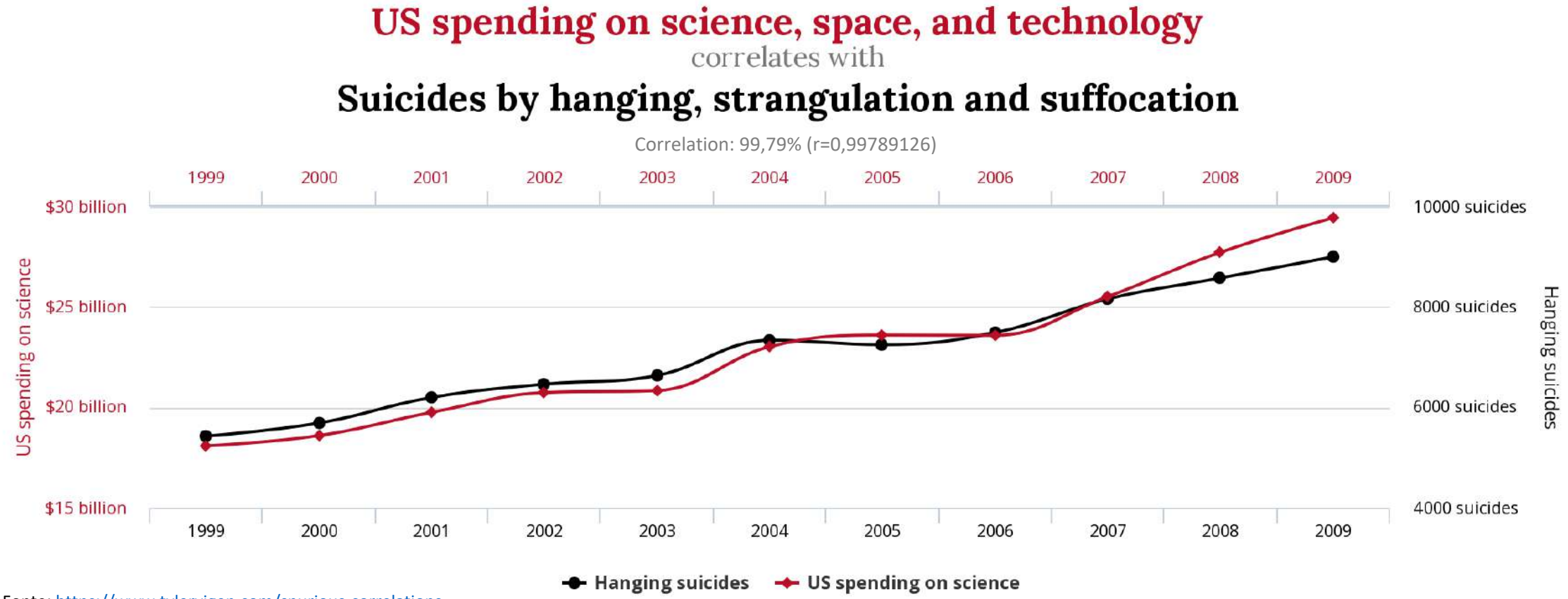
Vejamos alguns outros exemplos.

Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Exemplo 1: Gasto em Pesquisa no EUA vs. Suicídios



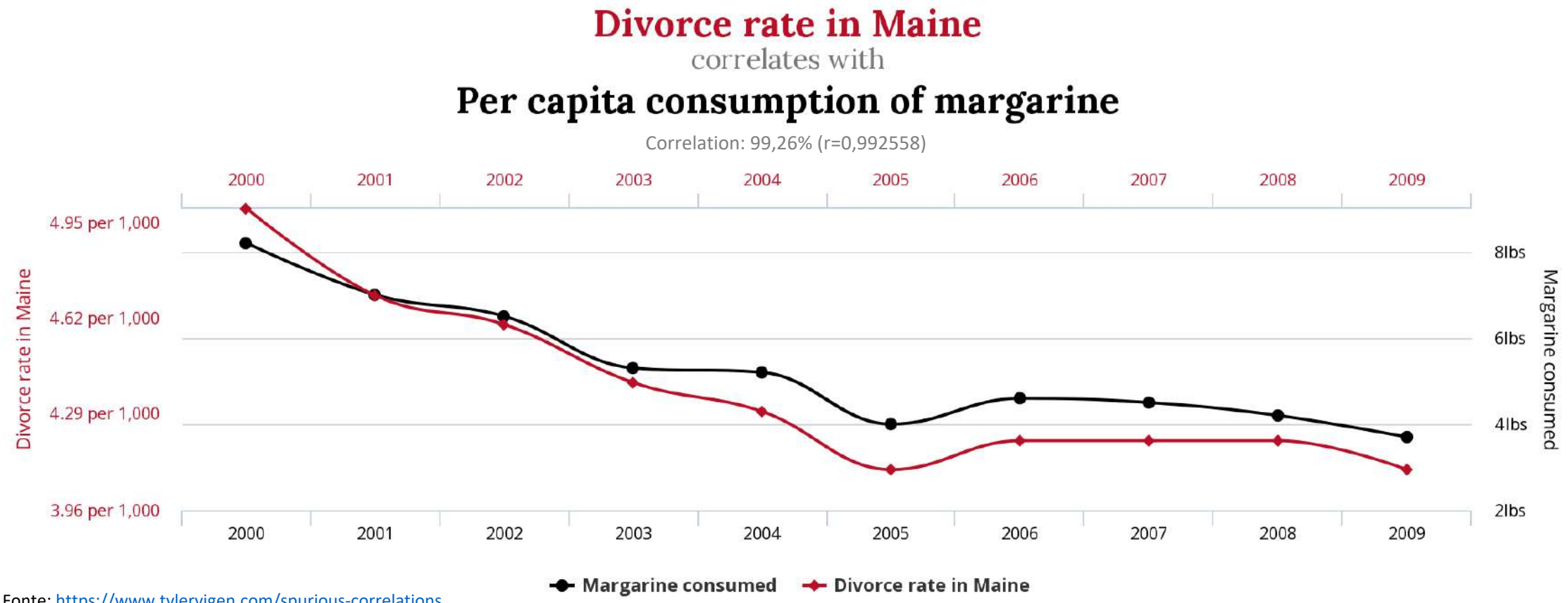
Fonte: https://www.tylervigen.com/spurious-correlations

Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Exemplo 2: Divórcios em Maine vs. Consumo de margarina



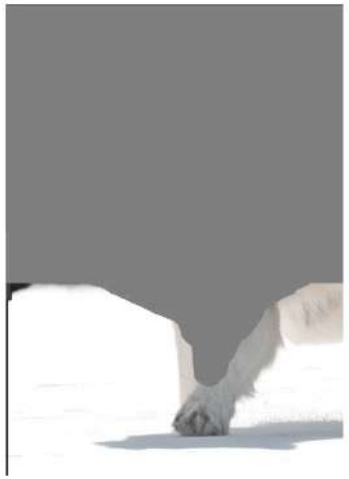
Fonte: https://www.tylervigen.com/spurious-correlations

Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Lobo



Lobo

Cachorro

Cachorro

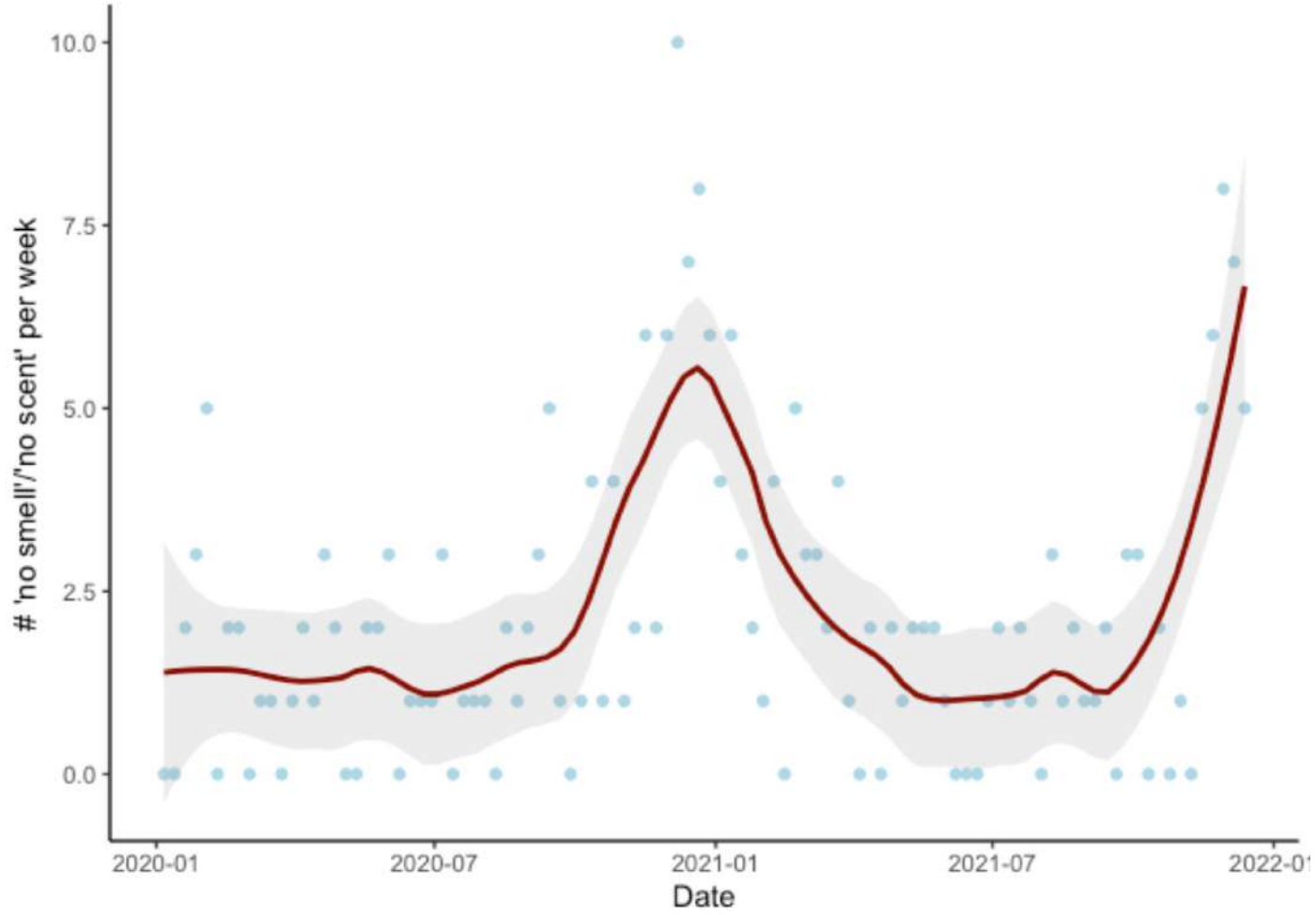
Fonte: https://www.analyticsvidhya.com/blog/2019/08/decoding-black-box-step-by-step-guide-interpretable-machine-learning-models-python/

Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Número de reclamações na Amazon para uma vela aromática (sem cheiro)

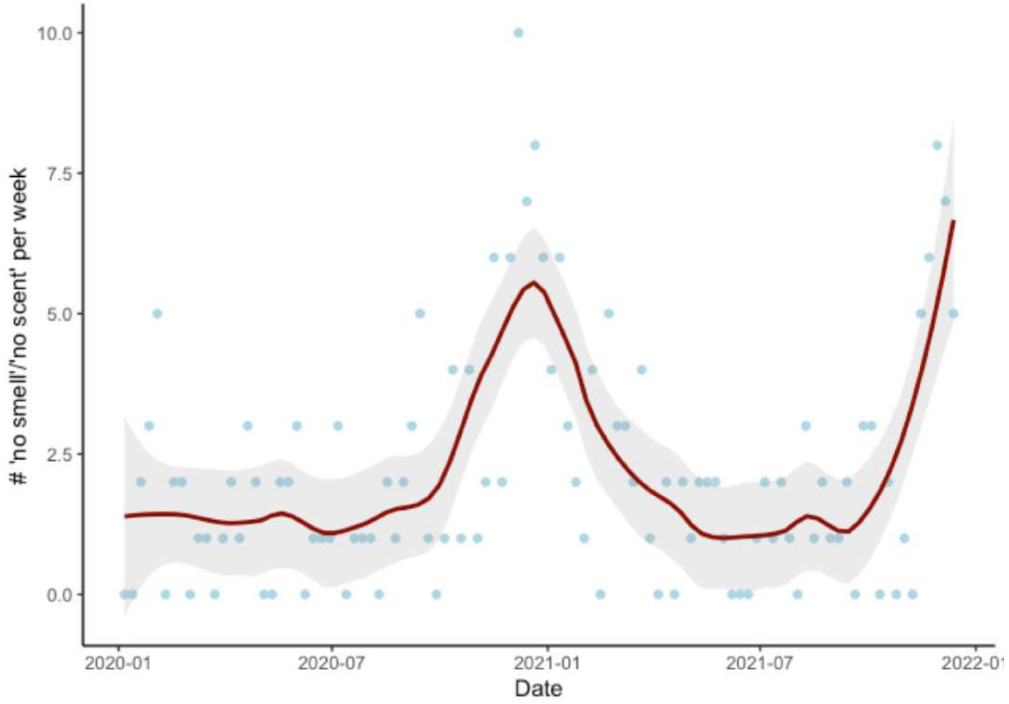


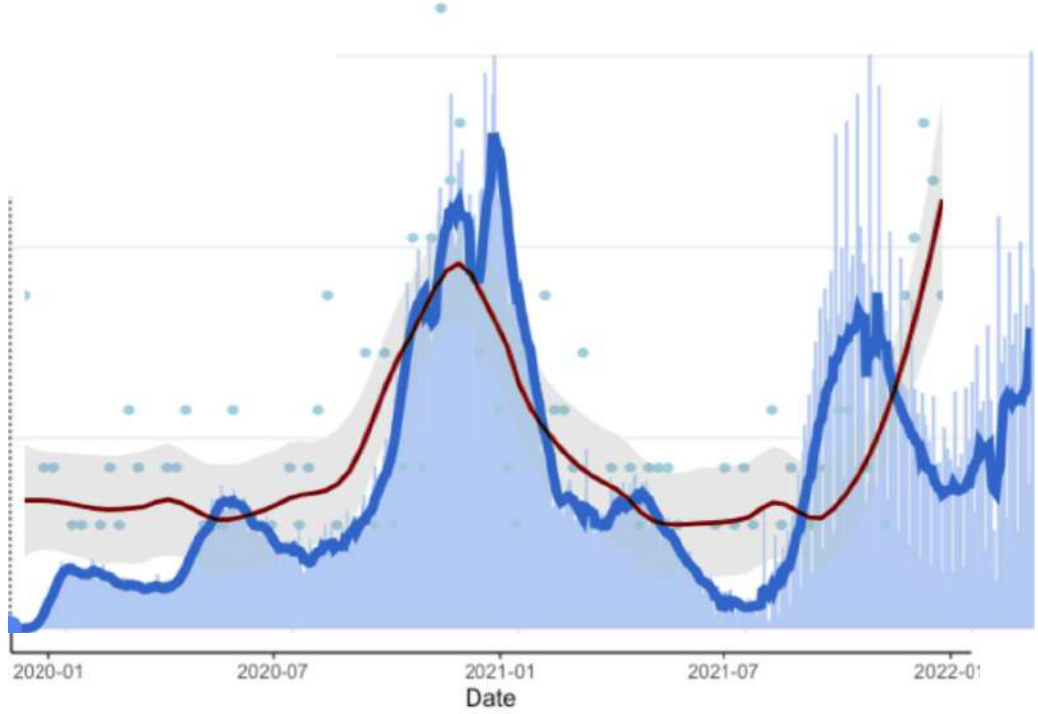
Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Número de reclamações na Amazon para uma vela aromática (sem cheiro)





Fonte: https://twitter.com/zornsllama/status/1473575508784955394

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade

Preditiva.ai

Fonte: https://www.cnbc.com/2022/10/18/negative-candle-reviews-and-covid-19-whats-the-correlation.html

Medidas e Gráficos

Exemplo 8 - Causalidade

Preditiva.ai

Conclusões:

- Se 2 variáveis estão **correlacionadas**, **pode ou não** haver **causalidade**
- Se houver **correlação** e não houver **causalidade** entre essas 2 variáveis, possivelmente há uma **3ª variável que não foi observada**
- **Mantenha-se cético**: busque **fortes evidências** para assumir a **causalidade**
- Antes de assumir a **causalidade** responda as seguintes perguntas:
  - **Por que** a variável A **causa** a variável B?
  - **Como** a variável A **causa** a variável B?

O mapa da CAGADA em Dados!

Parte 1

Preditiva.ai

Medidas e Gráficos

Medidas estatísticas

Unidades da base

Medidas vs Gráficos

Correlações

Causalidade

Projeções

Frequências imprecisas

Probabilidade e Possibilidade

Amostragem

Viés de seleção

Erros de Coleta

Tamanho de Amostra

Essas cagadas ficam para a parte 2 ... rs

Resumindo os erros da live (Parte 1)

Guarda esse check-list

Preditiva.ai

	Cagada 💩	Como evitar?
1	Usar apenas uma medida para resumir os dados.	Quanto mais medidas resumo usar, melhor conseguirá entender os dados analisados.
2	Assumir a unidade da base utilizada sem perguntar para o analista gerador da análise.	Deixar claro qual unidade da base deve ser trabalhada (tanto se você for o cliente do trabalho quanto você for o criador da análise).
3	As vezes nem muitas medidas resumo são suficientes para entender seus dados.	Sempre que possível plotar os dados em um gráfico analítico e comparar com as medidas resumo.
4	A correlação entre seus dados pode mudar ao agrupá-los de formas diferentes. (Paradoxo de Simpson)	Sempre plote o gráfico para cada grupo mais importante de sua base de dados.
5	Assumir que uma variável causa a outra só porque as variáveis estão correlacionadas. Correlação não significa causalidade.	Mantenha-se cético: busque fortes evidências para assumir a causalidade. Se não estiver confiante da causalidade, evite usar essa correlação.

Você quer **acelerar** seu desenvolvimento e se sentir **ainda mais confiante** para resolver problemas como mostrei aqui ?

Como **complemento** à nossa formação principal “**Gerando Valor com Dados**”, criamos o programa:

Preditiva.ai

ACELERA

Evolua 6 meses de experiência em Dados em apenas 1 mês

Como funciona?

ACELERA

Evolua 6 meses de experiência em dados em apenas 1 mês

O Acelera é um **programa ao vivo** onde você vai:

1. **Resolver problemas reais** utilizando as **técnicas**, **ferramentas** e **metodologias** aprendidas na Plataforma Preditiva com o **apoio de mentores experientes**.

2. Ter a experiência de trabalhar em um **projeto de dados colaborativo,** com colegas que têm o mesmo objetivo que você e o **apoio de mentores experientes**.

3. Treinar seu **Data Storytelling** **apresentando suas análises** para os gestores de diversas áreas e obtendo feedback em tempo real.

4. Aprender em **apenas 1 mês**, de forma **100% prática** a extrair ***insights* dos dados**, e estar pronto para buscar sua transição de carreira.

Preditiva.ai

ACELERA

Evolua 6 meses de experiência em dados em apenas 1 mês

Carga Horária: 15h
Serão 5 encontros ao vivo das 19h30 às 22h30 (3h cada encontro)

Certificado
Participando de **pelo menos 4 encontros**, você conquistará **seu certificado em Projetos de Dados**

Calendário

- **Dia 1**: 03-Abr-23: Onboarding, Team Building e Início do projeto
- **Dia 2**: 10-Abr-23: Desenvolvimento das Análises - Parte 1
- **Dia 3**: 17-Abr-23: Desenvolvimento das Análises - Parte 2
- **Dia 4**: 24-Abr-23: Avaliação dos Resultados e Preparação da Apresentação
- **Dia 5**: 08-Mai-23: Apresentação Final

Inscreva-se agora

Inscrições até 28/02

Preditiva.ai