INTRODUÇÃO À CIÊNCIA DE DADOS

Machine Learning
Ética e Privacidade

MACHINE LEARNING

- Relembrando, o nosso objetivo é usar dados para desenvolver modelos capazes de prever:
  - Se um email é spam ou não
  - Se uma transação de crédito é fraudulenta ou não
  - Se um cliente está a ponto de solicitar o cancelamento de um serviço
  - Ajustar um anúncio com maior capacidade de ser clicado por um cliente

MACHINE LEARNING

- Embora seja possível criar muitas regras que funcionem bem com dados de treinamento, é importante que as regras operem bem com novos dados, ou seja, os conjuntos de testes. Isso que se espera de uma predição!
- Podem haver discrepâncias entre o modelo treinado e os testes que são feitos com a outra parte dos dados.
- Espera-se sempre que não aconteça no modelo nem o overfitting (sobreajuste) nem o underfitting (subajuste). A busca do cientista de dados é sempre pelo modelo mais equilibrado possível.

MACHINE LEARNING

- Quanto mais complexo o modelo for, melhor preverá os dados de treinamento. No entanto, se o modelo se tornar muito complexo, concentrar-se-á demais em cada ponto de dados individual no conjunto de treinamento, e o modelo não será generalizado para novos dados.

- Há um ponto ideal no meio que renderá o melhor desempenho de generalização, este é o modelo que deve ser buscado e encontrado.

OVERFITTING (SOBREAJUSTE)

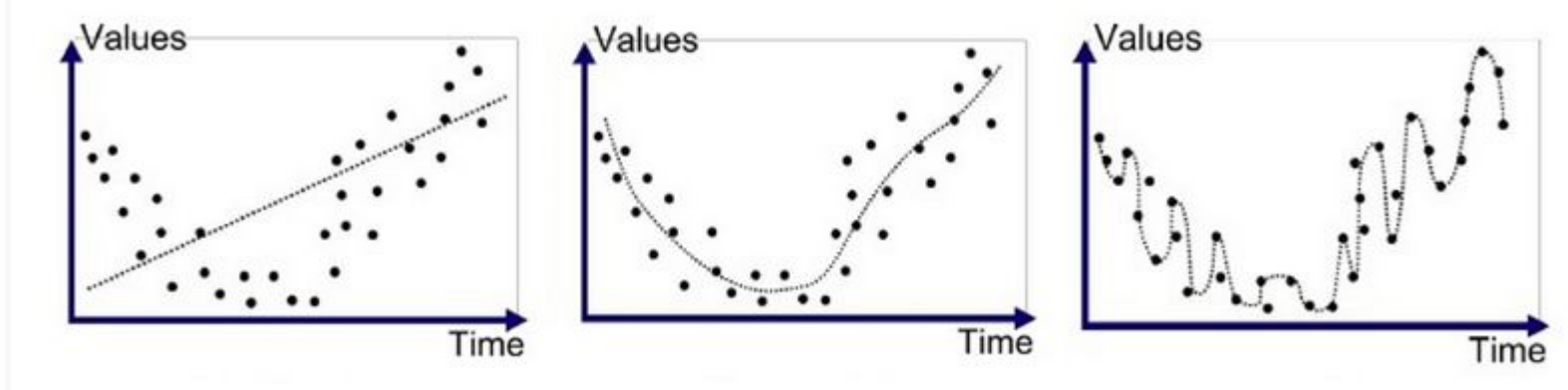
- Ocorre quando se ajusta um modelo muito de perto às particularidades do conjunto de treinamento e obtém um modelo que funciona bem no conjunto de treinamento, mas não é capaz de generalizar para novos dados.

UNDERFITTING (SUBAJUSTE)

- Se o modelo é muito simples, então pode não ser capaz de capturar todos os aspectos e variabilidade dos dados, e o modelo se sairá mal, mesmo no conjunto de treinamento. Escolher um modelo simples demais é chamado de underfitting.

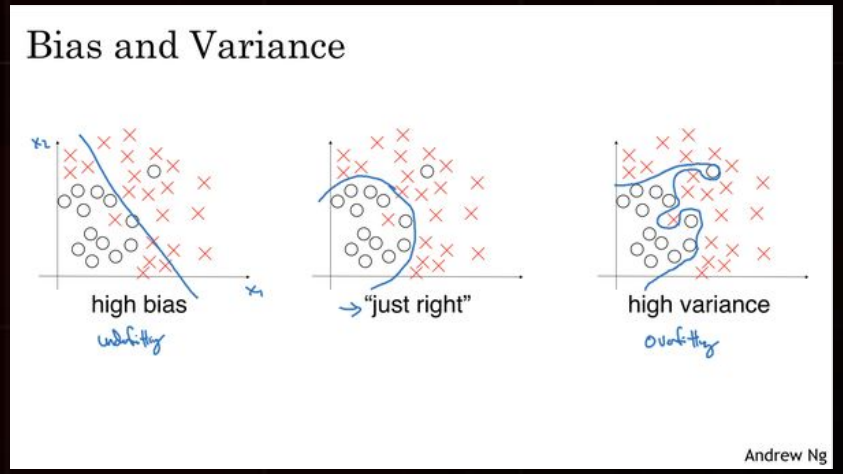
UNDERFITTING - OVERFITTING

- A primeira imagem representa um underfitting, modelo não consegue nem atender ao treinamento.
- A terceira imagem é um overfitting, o modelo está extremamente condicionado ao treinamento.
- A imagem do meio é um bom modelo ajustado que atende ao treino e deve ser bom nos dados de teste!!



VIÉS E VARIÂNCIA

- O viés (bias em inglês) é a incapacidade de um modelo em estabelecer a relação entre as variáveis e o objeto a ser predito.
- A variância está relacionada à sensibilidade de um modelo. Se um modelo é muito sensível ao treinamento, ele não é capaz de acertar os resultados nos dados de teste.



REGULARIZAÇÃO

- Regularização é uma das técnicas utilizadas para minimizar os problemas de variância.
- É um processo de penalização nos dados, de forma a minimizar o erro da generalização do modelo sem afetar muito o modelo baseado no treinamento.
- Há várias estratégias para efetivar regularizações, entretanto o mais importante por enquanto é saber que é muito comum (e veremos isso mais a frente) a utilização desse recurso para realizar ajustes nos modelos.

PRIVACIDADE

- Ao tratar dados de uma forma geral é muito importante observarmos que após inúmeros caso de exposição de dados pessoais, alguns países, inclusive o Brasil, vêm se preocupando com a questão de como lidar com dados pessoais.
- No Brasil temos a LGPD:
  - A Lei Geral de Proteção de Dados Pessoais (LGPD ou LGPDP), Lei nº 13.709/2018[1], é a legislação brasileira que regula as atividades de tratamento de dados pessoais e que também altera os artigos 7º e 16 do Marco Civil da Internet.

PRIVACIDADE

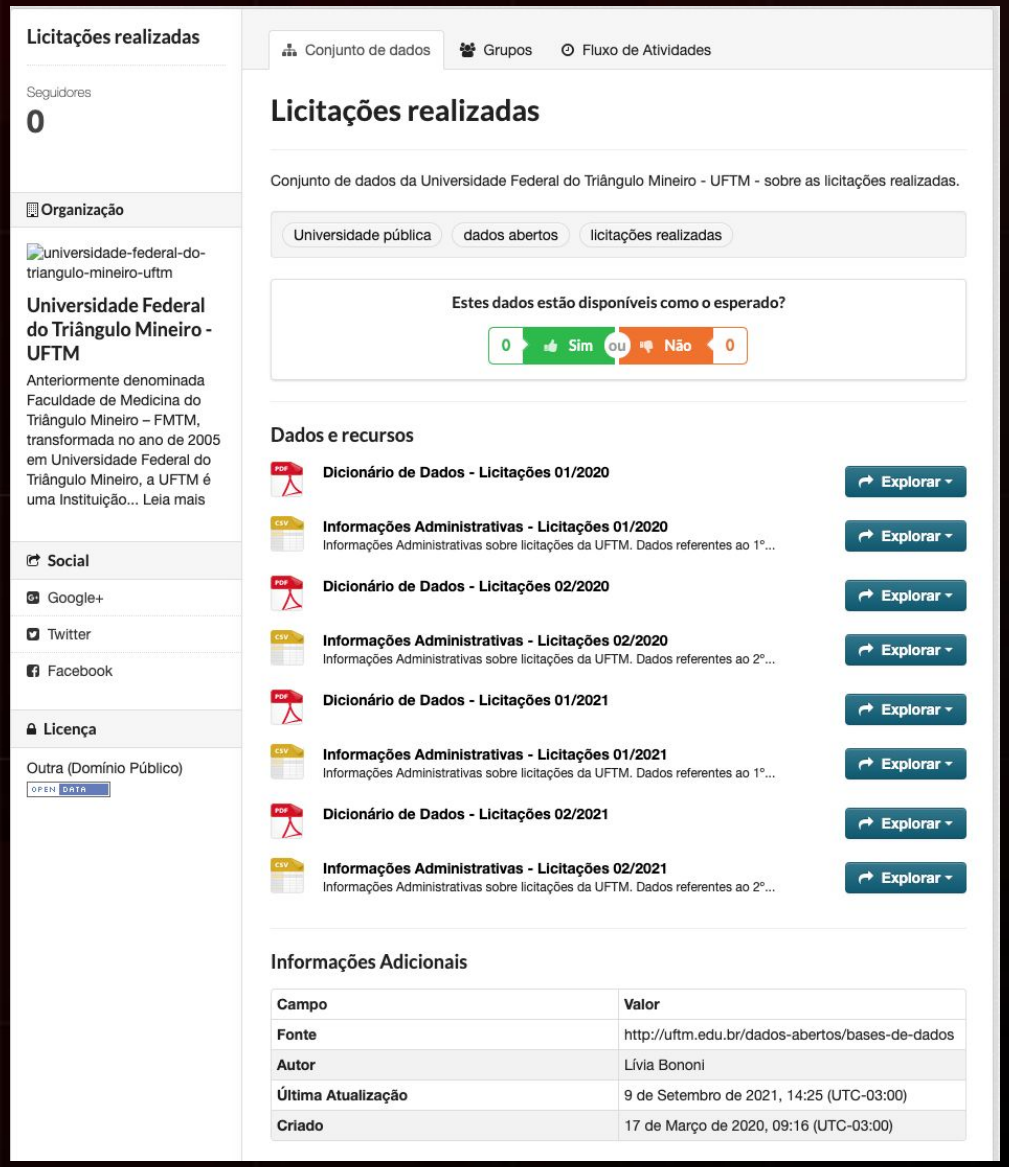
- Ressalta-se a necessidade do cuidado e de conhecer a legislação para evitar lidar com dados não públicos.
- Algumas técnicas de anonimização de dados pessoais são recorrentes, inclusive há algoritmos para tal.
- O objetivo aqui não é discutir as técnicas, mas sim destacar a necessidade da compreensão e bom senso ao lidar, principalmente, com dados pessoais, mas também com dados privados de um modo geral.

PRIVACIDADE - DICAS

- Algumas dicas importantes ao lidar com dados de um modo geral:
  - Esses dados têm algum tipo de licença pública?
  - Qual é a origem desses dados, ou seja, qual é a proveniência deles.
  - Além da proveniência (de onde vieram), como eles chegaram até você? Foram comprados, doados, o processo de aquisição é legal?
  - Esses dados estão atualizados?

PRIVACIDADE - DICAS

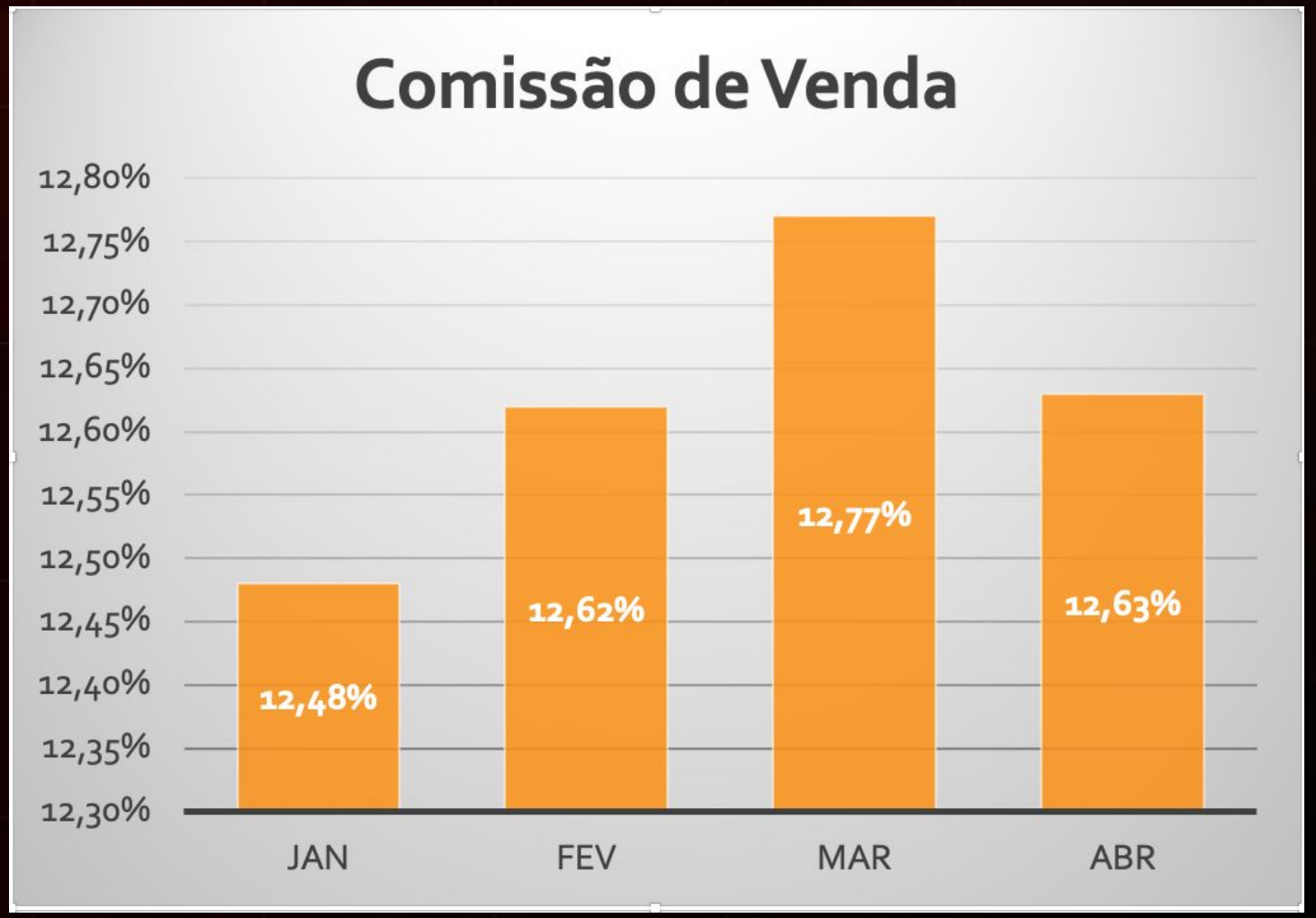
- Observe:



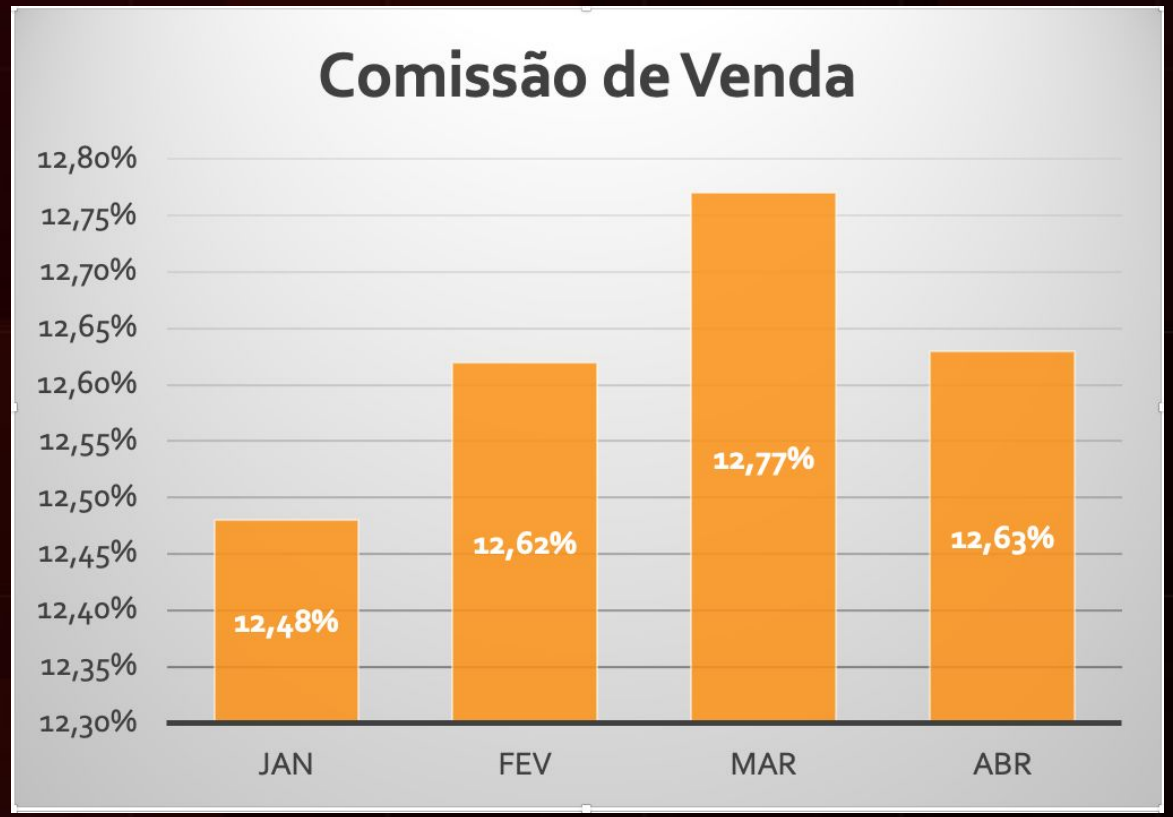
DISCRIMINAÇÃO

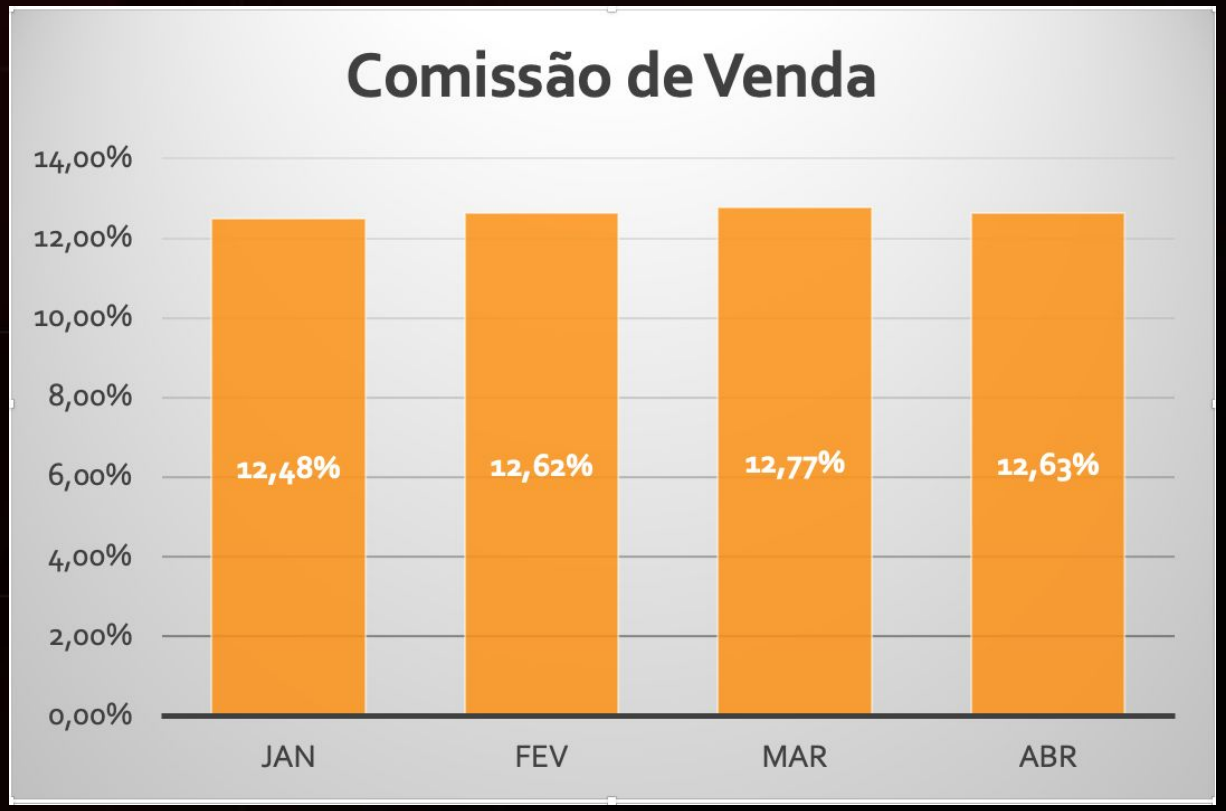
- Um fator pouco discutido ao lidar com dados é a questão ética e discriminatória das análises.
- A questão da discriminação deve ser sempre considerada para evitar que grupos minoritários sejam penalizados por dados que sempre privilegiem grupos majoritários.
  - Exemplo: buracos em Boston (estudo de 2013).
  - Exemplo: estudo sobre saúde em um país onde 99% da população usa o mesmo sistema, ou analisar a saúde baseado em um país onde 30% da população usa um determinado sistema (público ou privado)

ÉTICA (VEJA O GRÁFICO)



ÉTICA (OBSERVE NOVAMENTE)



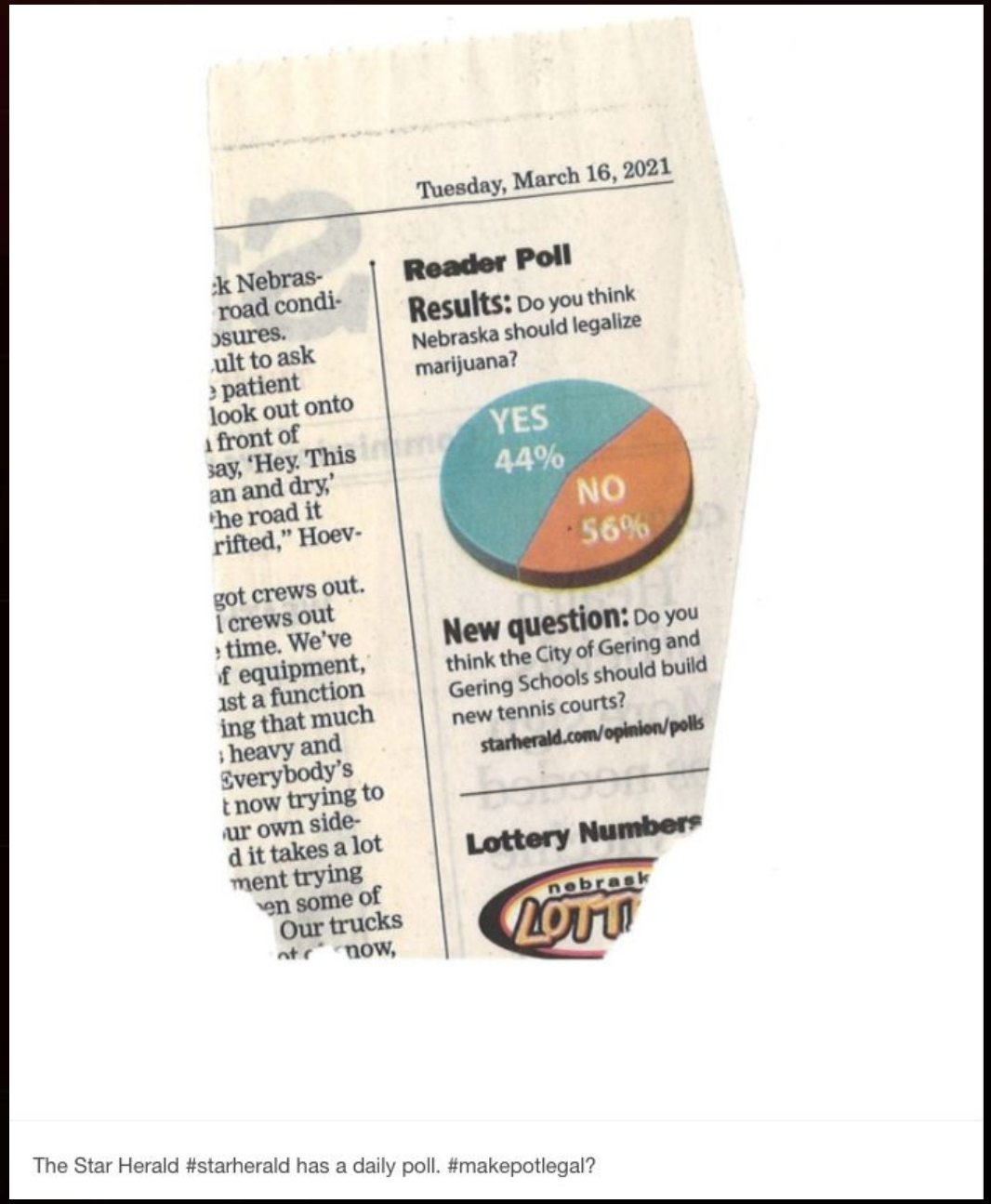


ÉTICA

- Onde está o erro dos gráficos apresentados anteriormente?
- É um problema ético, dos dados, da fonte, de quem fez a análise, de como fez a análise?
- Um problema ético está presente quando se deseja demonstrar a falta de realidade, ou apresentar resultados que não condizem com os dados de treinamento.
- Podemos entender como tendenciosidade, manipulação, fraude, inocência, falta de conhecimento?

ÉTICA

- Podemos ter aqui apenas um problema de visualização?
- Veja mais em: https://viz.wtf/



ÉTICA

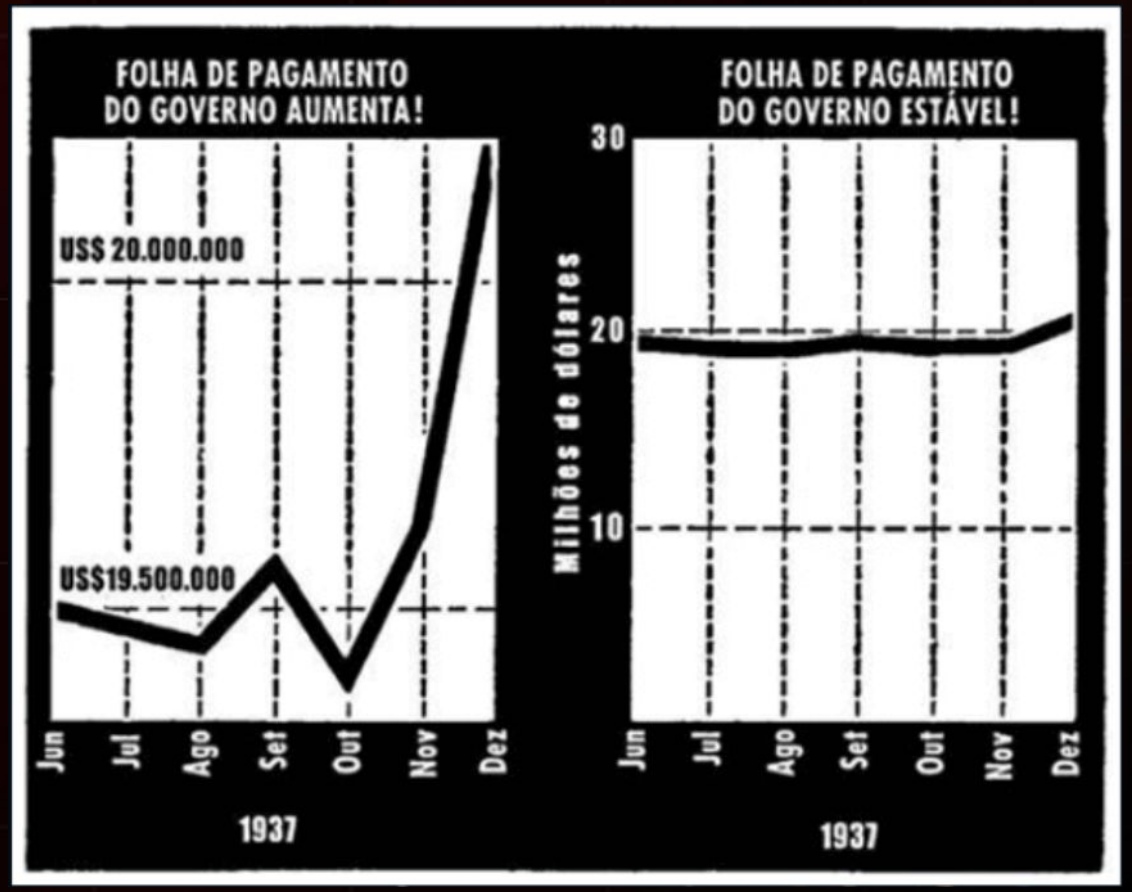
- Usar dados (ou falsas medidas) quando é conveniente.

Nro Func	Salário
1	930
2	950
3	990
4	1000
5	1090
6	1100
7	1115
8	1200
9	1200
10	1250
11	40000
Média	4620,45
Mediana	1100
Moda	1200

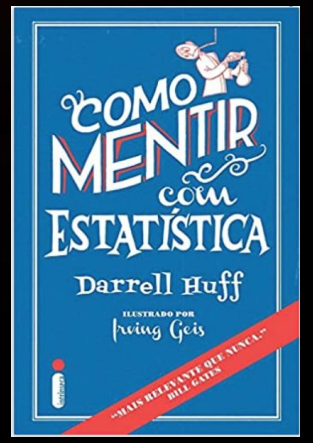
Nro Func	Salário
1	930
2	950
3	990
4	1000
5	1090
6	1100
7	1115
8	1200
9	1200
10	1250
11	
Média	1082,50
Mediana	1095
Moda	1200

ÉTICA

- Apelar nas visualizações!



Fonte:



FINALIZANDO

- Muito comum que os primeiros modelos tenham problemas de overfitting ou underfitting.
- Em muitos casos vamos precisar de técnicas de ajuste para os modelos.
- As questões de ética e privacidade são de fundamental importância, desde a coleta dos dados até a entrega dos resultados, geralmente por meio de visualizações.

REFERÊNCIAS

- Parte do conteúdo dessa aula é baseado no livro:

Grus, Joel. Data Science do Zero. Disponível em: Minha Biblioteca, (2nd edição). Editora Alta Books, 2021.

INTRODUÇÃO À CIÊNCIA DE DADOS

Machine Learning
Ética e Privacidade