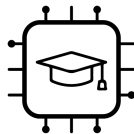


# **CLUBE DE ASSINATURAS**

# **UNIVERSIDADE DOS DADOS**

Material da Semana 3 (20/02/2023)



**Universidade dos Dados**

## 0. INTRODUÇÃO

Olá, turma!

Sejam bem-vindos a mais uma semana do Clube de Assinaturas! Dessa vez, o material está dividido em 3 partes: entrevistas de emprego, regressão linear e análise de dados na prática. A ideia é que vocês ganhem cada vez mais conhecimentos para se destacarem de diversas formas, sabendo lidar com entrevistas, entendendo o que está por trás do fit-predict do Scikit-Learn e ganhando insights para criarem suas próprias análises. Essa última é minha seção preferida de todos os materiais, penso que será *game changer* para muitos aqui.

Dessa vez, fiz questão de já deixar no começo do material quais os pré-requisitos, pois a seção 2 será muito complexa para quem nunca teve aula de Cálculo. Aliás, se você não tem o conhecimento necessário para essa seção, não se desespere, pois (1) você ainda pode usar o algoritmo numa boa e eu vou te ajudar com isso nas próximas semanas; e (2) você já entende como o algoritmo funciona, portanto, não vai passar vergonha nas entrevistas. Na verdade, é provável que você saiba mais do que a maioria dos candidatos. Se quiser eventualmente aprender essa parte mais matemática, sugiro assistir às aulas de Cálculo da Khan Academy ou pegar o livro, gratuito, Mathematics for Machine Learning e estudar os capítulos que falam de derivada parcial. Também vale olhar cursos baratinhos que sejam focados nisso.

A maioria dos alunos não tem conseguido estudar todo o material que eu venho passando, pois estavam bem mais extensos do que o combinado no lançamento do Clube. Por esse motivo, achei que fazer um material um pouco mais focado essa semana não prejudicaria ninguém. Sempre há atividades extras sendo sugeridas aqui e no grupo, além de estudos complementares que já ensinei como fazer. Então não acho que alguém venha a se sentir lesado ou vá ficar ocioso.

Sobre os pré-requisitos de cada seção:

1. ENTREVISTA DE MACHINE LEARNING: Esta seção será melhor aproveitada por pessoas que já estejam estudando Machine Learning. Com alguma ajuda da internet, é perfeitamente possível para qualquer um passar por essa seção, mas pode ser que ainda não seja o momento adequado para isso. Recomendo que salvem o material para ler futuramente, quando estiverem mais confortáveis com os temas.

2. REGRESSÃO LINEAR: Esta seção será melhor aproveitada por pessoas que tenham conhecimento em Cálculo (derivada) e Álgebra Linear (matrizes). Novamente, se este não for seu caso, guarde o conteúdo para o futuro.
3. DS NA PRÁTICA: Sem pré-requisito.
4. FALÁCIAS EM DADOS: Sem pré-requisito.

## 1. ENTREVISTA DE MACHINE LEARNING

### "COMO EU RESPONDERIA..."

Bom, como muitos de vocês estão passando pela fase dos processos seletivos, resolvi dar uma mãozinha contando como eu responderia algumas das questões mais comuns em entrevistas de emprego (e os tipos de respostas que eu espero de um candidato).

São raras as entrevistas em que você precisa demonstrar algo matematicamente, eu mesmo, em quase 10 anos de carreira, nunca passei por isso. Claro, talvez atualmente até exista algum desconto por já demonstrar o que sei publicamente todos os dias nas redes sociais, mas acho que o principal é que ninguém cobra isso mesmo. Normalmente, você terá que dar algumas respostas técnicas usando só da fala, pois isso demonstra sua capacidade de transpor um conhecimento técnico para a linguagem do dia a dia. Mas fiquem tranquilos que nem por isso a gente vai ignorar as questões matemáticas. Elas são sim úteis e a gente vai ver isso ao longo das próximas semanas - hoje mesmo já temos algumas demonstrações da regressão linear!

Bom, bora para as perguntas e respectivas respostas...

P: O que é regularização?

O que eu responderia: A regularização em machine learning é uma técnica que a gente usa para evitar o overfitting, evitar que o modelo se ajuste demais aos dados de treino e não consiga fazer previsões com novos dados (quando ele entrar em produção).

De forma simples e direta, o que o modelo de Machine Learning faz é encontrar padrões nos dados históricos para aí tentar fazer previsões quando surgirem novos dados. O problema é que às vezes o modelo é tão complexo e tenta fazer previsões tão precisas que ele se ajusta

quase que perfeitamente aos dados de treino, considerando no seu aprendizado até mesmo os ruídos. Seria como se o seu modelo decorasse os dados de treinamento, mas não realmente aprendesse o comportamento médio daquelas observações. Isso é bastante comum em modelos deep learning mais complexos. Aí quando isso acontece, seu modelo não vai conseguir fazer boas previsões em novos dados, ele não consegue generalizar o aprendizado e temos um problema. Para resolver este problema, aparece a regularização.

A regularização é uma forma de controlar a complexidade do modelo, ela adiciona um termo de penalidade à função de custo usada durante o treinamento. Então imagina assim, o seu modelo de Machine Learning busca minimizar ao máximo a função custo e pode até mesmo minimizar demais, caindo no overfitting, que acabamos de explicar. Essa penalidade vai evitar que essa minimização passe do ponto "saudável" e vai meio que corrigir o aprendizado. Esse termo adicionado pela regularização vai penalizar coeficientes grandes ou desnecessários, fazendo com que o modelo tenha soluções mais simples e assim realmente aprenda o comportamento dos seus dados, não apenas decorando algo que não vai ser extrapolado. Existem diferentes tipos de regularização, algumas bem comuns e que usamos para corrigir inclusive a regressão linear, um dos algoritmos mais populares em Machine Learning é a L1 (lasso), L2 (ridge) e Elastic Net.

Em suma, a regularização vai tentar evitar que um modelo se ajuste demais aos dados de treinamento, melhorando sua capacidade de generalização para novos dados.

P: O que são modelos de ensemble? Como você explicaria isso para um leigo?

O que eu responderia: Os modelos ensemble em machine learning são modelos que combinam diversas técnicas para obter uma previsão melhor. Ou seja, é uma tentativa de usar técnicas que sozinhas não seriam suficientes para se obter uma boa acurácia (ou qualquer métrica de avaliação), mas que quando combinadas, obtemos um modelo muito mais robusto.

Uma Random Forest é um exemplo de modelo de ensemble. O que ela faz é combinar as previsões de diversas árvores de decisão afim de obter um modelo preditivo que tenha uma previsão mais precisa.

Mas vamos evitar o tecnicismo aqui para entender a lógica. Pense assim, imagine um problema em que você quer obter o preço de aluguel ideal para um imóvel. Aí você tem um corretor que

pode estimar um certo preço a partir dos dados e de sua experiência. Este preço pode até ser bom, mas você poderia talvez conseguir algo melhor pegando a estimativa de outros 30 corretores com diferentes experiências e métodos. É bem provável que pegando a média de vários corretores você chegue mais próximo do preço correto.

Existem várias técnicas de modelos ensemble, como bagging e boosting. Cada técnica tem sua própria abordagem para combinar os modelos e gerar uma previsão final mais precisa. Você pode pegar várias previsões e aí fazer a previsão baseado na média ou na maioria, que seria o caso do bagging. Ou pegar uma predição e imputar em outro modelo, e aí um modelo aprenderia do anterior. Esse seria o caso do boosting.

P: Explique para um leigo o funcionamento de algum algoritmo de Machine Learning.

O que eu responderia: Bom, vamos falar da árvore de decisão, porque é uma estrutura que muita gente já acaba lidando no dia a dia, mesmo que às vezes de forma intuitiva.

A árvore de decisão é um algoritmo que funciona como se fosse um fluxograma de perguntas que levariam até uma resposta final, até uma tomada de decisão. Então você imagina que a gente tenha uma árvore para um modelo de risco de crédito. Na hora que um cliente for passar pelo modelo, ele vai cair numa série de perguntas nesse fluxograma, coisas como "número de imóveis próprios", "salário abaixo ou acima de determinada faixa", "se possui uma dívida em aberto ou não". Cada uma das possíveis respostas vai levar a uma decisão final do modelo, entre classificar a pessoa como boa ou má pagadora.

A escolha entre começar com a pergunta da dívida, ou do salário, a sequência de perguntas, vai ser determinada através da relevância da feature. E o algoritmo vai considerar a feature mais relevante aquela que consegue separar melhor o público. Para simplificar, imagine que a gente tenha a pergunta se a pessoa tem ou não imóvel e a pergunta se a pessoa tem ou não dívida em aberto. A dos imóveis, deixa um lado com metade sendo mau pagador e metade sendo bom pagador. A de dívida em aberto, deixaria todos os maus pagadores de um lado e os bons de outro, ou que seja todos os maus de um lado e o outro lado com 90% de bons pagadores e 10% de maus pagadores. Bom, então a pergunta da dívida separa melhor os bons dos maus clientes, ela é a que iria primeiro. Claro, isso simplificando. Na prática, temos duas possíveis métricas, o índice de Gini ou a Entropia.

P: O que é a loss function?

O que eu responderia: Uma loss function, uma função de perda, é o que o modelo de machine learning utiliza para saber se ele está indo bem ou não, se as previsões dele estão dentro do que a gente esperava, ou se estão melhores que a de outro modelo. É basicamente o que ele usa para saber o quão bem ele está conseguindo prever a saída correta a partir dos dados que foram imputados. É uma forma de quantificar o erro do modelo.

Quando a gente pede, entre aspas, para o modelo fazer a melhor previsão possível a partir dos dados históricos que temos, o que ele vai fazer é tentar encontrar o padrão que minimize a função perda. Pense na regressão linear, por exemplo. A gente passa os dados e o modelo vai tentar encontrar coeficientes que formam uma equação linear. E esses coeficientes vão ser encontrados olhando para a função perda, tentando sempre minimizá-la. Se você usar o MSE, o erro quadrático médio, ele vai pegar a somatória do quadrado da diferença entre o valor real e o valor estimado por aquela equação e dividir isso pelo tamanho do histórico, pelo número de observações. Essa função é o que chamamos de função perda.

O objetivo do modelo é minimizar o valor da função de perda, o que significa que ele está tentando fazer com que as saídas previstas sejam o mais próximas possível das saídas reais. Para isso, o modelo ajusta os pesos dos seus parâmetros durante o processo de treinamento, com base na função de perda.

Existem várias funções de perda diferentes, dependendo do tipo de problema que o modelo está tentando resolver. Por exemplo, uma função de perda comum para problemas de classificação binária é a binary cross-entropy, enquanto uma função de perda comum para problemas de regressão é o mean squared error que acabamos de descrever.

Agora, deixe você suas respostas, com base no que aprendeu na semana passada:

**P: O que é um modelo de machine learning?**

**P: O que é uma regressão linear?**

**P: Explique o que é overfitting para um diretor leigo em Machine Learning.**

## **2. REGRESSÃO LINEAR**

Enfim, vamos aos primeiros conteúdos que vão se aprofundar na matemática de Machine Learning. É hora de você entender o que está por trás de uma regressão linear!

### **2.1 DERIVANDO OS PARÂMETROS**

Vocês já aprenderam no material da semana passada o que é uma regressão linear e qual a lógica por trás dela, certo?

Apenas reiterando, o que acontece é que buscamos uma equação linear para descrever a relação de  $N$  variáveis, chamadas de features, com uma outra variável, chamada de alvo. Essa equação linear tem o desenho daquela que aprendemos no colégio, na aula sobre equação da reta, e, para encontrá-la, a gente tenta diminuir o quanto nossa reta "erra" suas previsões. Se você não se recorda, volte àquela aula e retome os conceitos de regressão linear e loss function.

Agora, se você já estudou Cálculo, você talvez já suspeite como podemos encontrar os coeficientes dessa reta. Lembra que a gente busca minimizar uma função (*loss function*) na hora de encontrar nossa "melhor" reta, a que vai ter as melhores previsões, a que "erra" menos? Bom, pense bem, como a gente encontra o mínimo de uma função? **Pela derivada!**

Já que você já sabe a ideia por trás de encontrar os parâmetros - minimizando a função custo, que é onde você "erra menos" -, a demonstração de como chegar neles ficou uma mamata só! Vou deixar aqui algumas demonstrações, uma que achei legal porque revisa as explicações sobre regressão linear, de uma maneira mais formal; outra do meu blog, feita à mão; outra do Geeks for Geeks, com representações matriciais (pode facilitar a depender de como foi seu ensino superior) e uma do Free Code Camp, apenas como alternativa adicional.

Só um ponto antes de olharmos as demonstrações: essa é UMA das formas de se chegar no mínimo da função custo. A gente chama isso de "normal equation", mas você vai ver lá na

frente que também é possível chegar usando o famoso Gradiente Descendente. Não precisa se preocupar com este novo termo por enquanto, apenas saiba de sua existência.

Demonstração 1: <https://www.obaricentrodamente.com/2010/07/regressao-linear.html>

Demonstração 2:

<https://estatsite.com.br/2019/10/17/derivando-os-parametros-de-uma-regressao-linear-simples/>

Demonstração 3: <https://www.geeksforgeeks.org/ml-normal-equation-in-linear-regression/>

Demonstração 4:

<https://www.freecodecamp.org/portuguese/news/aprendizagem-da-maquina-uma-introducao-ao-erro-quadratico-medio-e-linhas-de-regressao/>

## 2.2 PREMISSAS, INFERÊNCIA E PREDIÇÃO

Quem faz parte do clube, provavelmente já me acompanha há um bom tempo e já ouviu minha eterna discussão sobre inferência x predição, premissas da regressão linear, multicolinearidade e por aí vai. Tenho certeza de que já ficou até cansativo, mas eu prometo que vai ficar interessante e que saber o conteúdo dessa seção vai FACILITAR sua vida, não o contrário.

Eu bato muito nessa tecla porque vocês vão ver que existem algumas premissas numa regressão linear que são utilizadas para garantir que aqueles coeficientes da equação são os coeficientes que chamamos de BLUE, Best Linear Unbiased Estimators (Melhores Estimadores Lineares Não-Viesados). Ou seja, elas só dizem respeito a encontrar esses parâmetros. O problema é que algumas pessoas ensinam elas como sendo todas necessárias independente da situação. Tanto é um uso meio genérico, sem entendimento da situação, que a gente vê muitas frases do tipo "quando essas suposições não são atendidas, os resultados de análise de regressão podem ser enganosos e o modelo pode não ter um bom desempenho". Enganosos como? Não tem bom desempenho em que sentido? Não faz boas predições? Os coeficientes não são reais? O que exatamente é afetado? Tudo?

Entende agora meu ponto?

Agora que vocês entendem que as premissas dizem respeito aos estimadores, que é preciso ir um pouco além na compreensão do problema e da motivação da regressão linear, podemos prosseguir...



Quero primeiro que vocês entendam quais são as premissas necessárias para obter os tais estimadores BLUE e depois quero que vocês vejam as demonstrações matemáticas que mostram como cada premissa afeta seu modelo. Vocês vão ver, por exemplo, que multicolinearidade não afeta a predição, apenas diz respeito, repetindo mais uma vez, aos estimadores.

Por fim, vou deixar uma discussão que eu quero muito que todos leiam, para entender a distinção entre multicolinearidade e correlação, algo que causa muita confusão por aí!

## **PREMISSAS DE UMA REGRESSÃO LINEAR**

Antes de entrar nos materiais, vocês vão notar que as premissas para se obter estimadores BLUE não são as mesmas a depender do site, o que pode confundir vocês. Não se preocupem, normalmente, isso ocorre porque algum autor desconsiderou alguma coisa, por ela poder ser inferida de outra já dada. Então, ao invés de colocar a premissa A e a B, ele coloca só a A, por entender que a B pode ser obtida a partir da A. Eu costumo considerar como premissas, as seguintes questões: linearidade nos parâmetros, observações independentes entre si, variância constante nos erros (homoscedasticidade), normalidade dos resíduos e multicolinearidade não-perfeita.

A seguir, vocês vão entender um pouco mais sobre elas!

## **PREMISSAS EM PYTHON**

Começamos observando como validar essas premissas usando Python. Foi difícil achar alguém que usasse VIF (Variance Inflation Factor) para avaliar multicolinearidade e não usasse a matriz de correlação - mais para frente, vocês entenderão a diferença entre correlação e multicolinearidade e entenderão porque é errado usar matriz de correlação para capturar multicolinearidade. Felizmente, com muita busca, encontrei o notebook ideal:

<https://www.kaggle.com/code/simranjain17/linear-regression-assumptions-code>

## **COMO AS PREMISSAS AFETAM O MODELO**

Estas demonstrações não são triviais, portanto, vocês podem tomar o tempo que for necessário para absorver o material. Se for preciso, divida-o em pedaços. O mais importante, vocês já


sabem, como o algoritmo funciona e o que está acontecendo por trás dele para encontrar os parâmetros. Também já sabe como as premissas são utilizadas e que é preciso cautela na hora de dizer que um modelo "não terá bom desempenho" ou que terá "resultados enganosos". O texto aqui é para quem quiser mergulhar na matemática das premissas:

<https://medium.com/towards-data-science/ols-linear-regression-gauss-markov-blue-and-understanding-the-math-453d7cc630a5>

## CORRELAÇÃO VS MULTICOLINEARIDADE

Essa confusão é uma das mais comuns na internet e em diversos materiais. Digite agora mesmo "multicolinearidade vs correlação" ou "colinearidade vs correlação" e você vai ver dezenas e dezenas de pessoas dizendo que alta correlação implica em colinearidade, ou o contrário. No nosso clube, isso não vai acontecer. Colinearidade é quando uma ou mais variáveis são capazes de prever outra, elas são colineares (pense geometricamente). Imagine um cenário onde 9 variáveis sejam capazes de prever uma outra. Cada uma dessas 9 representa um pedacinho dessa última, ou seja, elas não possuem alta correlação com ela. Entretanto, temos multicolinearidade perfeita!

Veja o Peter Flohm falando sobre a discussão:



**Peter** · Follow

Independent statistical consultant for researchers in behavioral, social and medical sciences · 3y





Related

**What level of correlation indicates multicollinearity?**

None. You cannot judge multicollinearity from correlation. The best method is to use condition indexes.

You can have very high collinearity with no correlation above 0.1.

826 views · View 5 upvotes

 5   1 

<https://qr.ae/prUb0z>



Peter · Follow

... X

Independent statistical consultant for researchers in behavioral, social and medical sciences · 3y

Related I read some answer that there can be scenarios where correlation is not significant but two variables can be multi collinear. How multicollinearity is different from correlation?

Collinearity can involve more than two variables. For an artificial example, imagine that there are 9 variables that are independent and a tenth variable that is the sum of the first 9. Now, all the correlations will be low (around 0.11) but there will be perfect collinearity.

301 views · View 5 upvotes · Answer requested by Payal Bhatia



5



1



<https://qr.ae/prUg1w>

**E aí, curtiu aprender mais da matemática do algoritmo?**

Se você conseguiu pegar pelo menos uma parte desta seção, saiba que você já tem um conhecimento raríssimo em nossa área. Se quiser se aventurar rodando alguma dessas demonstrações no Python e ir postando no Medium, pode ter certeza de que vai deixar seu portfólio MUITO valioso!

### 3. DATA SCIENCE NA PRÁTICA

Esta seção talvez seja a preferida da maioria, mesmo se considerarmos as de outras semanas. No vídeo a seguir, vou mostrar para vocês uma análise de dados, comentando como seria no mundo real, como vocês podem fazer quando estiverem sozinhos e como usar em seu portfólio! Bora?

**Análise de Dados na Prática (Parte 1):** <https://youtu.be/UtoiT7lj-sk>

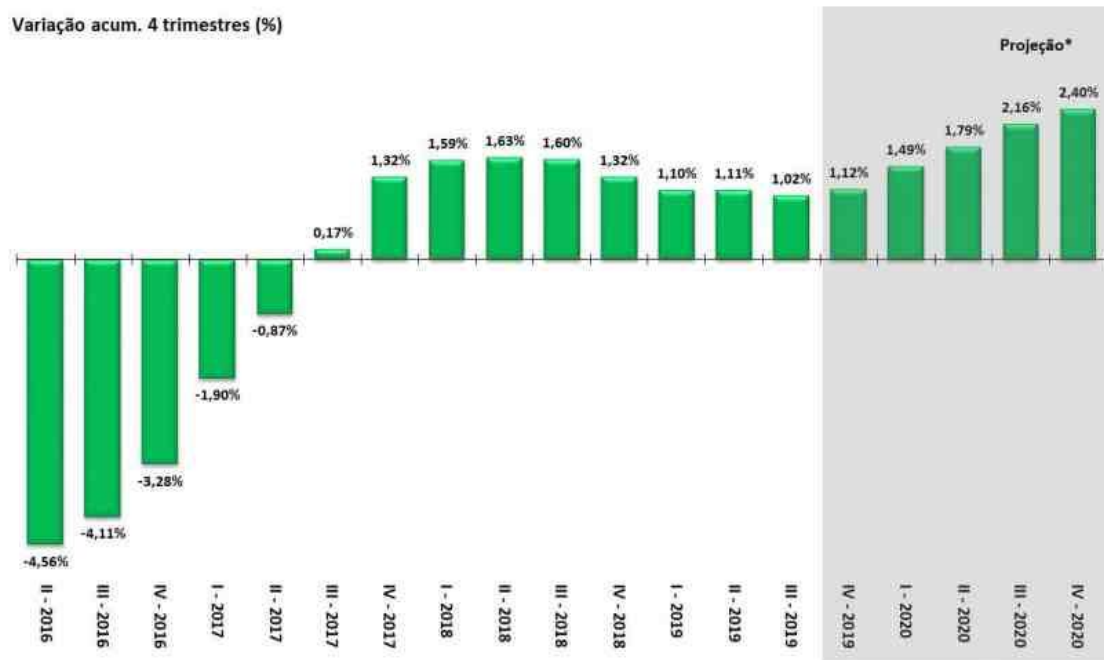
**Análise de Dados na Prática (Parte 2):** <https://youtu.be/5W4ukEbQ2Z4>

Como você viu nos vídeos acima, você tem um desafio para casa: criar análises para seu portfólio. Para facilitar ainda mais seu trabalho, aqui vão algumas dicas sobre como escolher o gráfico a ser utilizado na investigação, ou outro tipo de abordagem:

- Pense bem no storytelling e no público. O que você quer mostrar exatamente? Pretende mostrar, por exemplo, o que está causando a queda nas vendas da empresa? Que tal começar contextualizando o perfil dos clientes, mostrar a distribuição de gêneros e idade, depois mostrar numa linha do tempo quando começa a queda. E aí, vai chegando na conclusão, que mostra quais variáveis se relacionam com aquela queda.
- Algumas escolhas tradicionais de gráficos/análise:
  - Distribuição de variável contínua (ex.: idade, salário, altura, preço): boxplot, histograma e swarmplot.
  - Distribuição de variável categórica (classe social, gênero, país, profissão): mostre com gráficos de barras qual a frequência de cada categoria da variável em questão.
  - Relação entre duas variáveis contínuas: Gráfico de dispersão.
  - Relação entre uma variável categórica e uma contínua: Utilize boxplots para as diferentes categorias. Ou seja, coloque no eixo x a variável categórica e no y a variável contínua.
  - Relação entre duas categóricas: Uma tabela cruzada com a contagem seria interessante. Você pode incluir o percentual da linha ou da coluna.
- Se for incluir o ponto anterior num slide, pode usar de marcações com círculos e legendas em locais importantes. Veja as figuras 1, 2 e 3.
- **Evite:** gráficos de pizza e de rosca, 2 eixos-y.

## Evolução do PIB e Projeções Trimestrais

ATIVIDADE ECONÔMICA



Fonte: IBGE, Contas Nacionais Trimestrais. \*Projeção: Grade de Parâmetros Macroeconômicos, de janeiro/2020.

Fig 1: <https://www.ocafezinho.com/2020/01/14/os-numeros-da-producao-brasileira-segundo-o-ministerio-da-economia/>

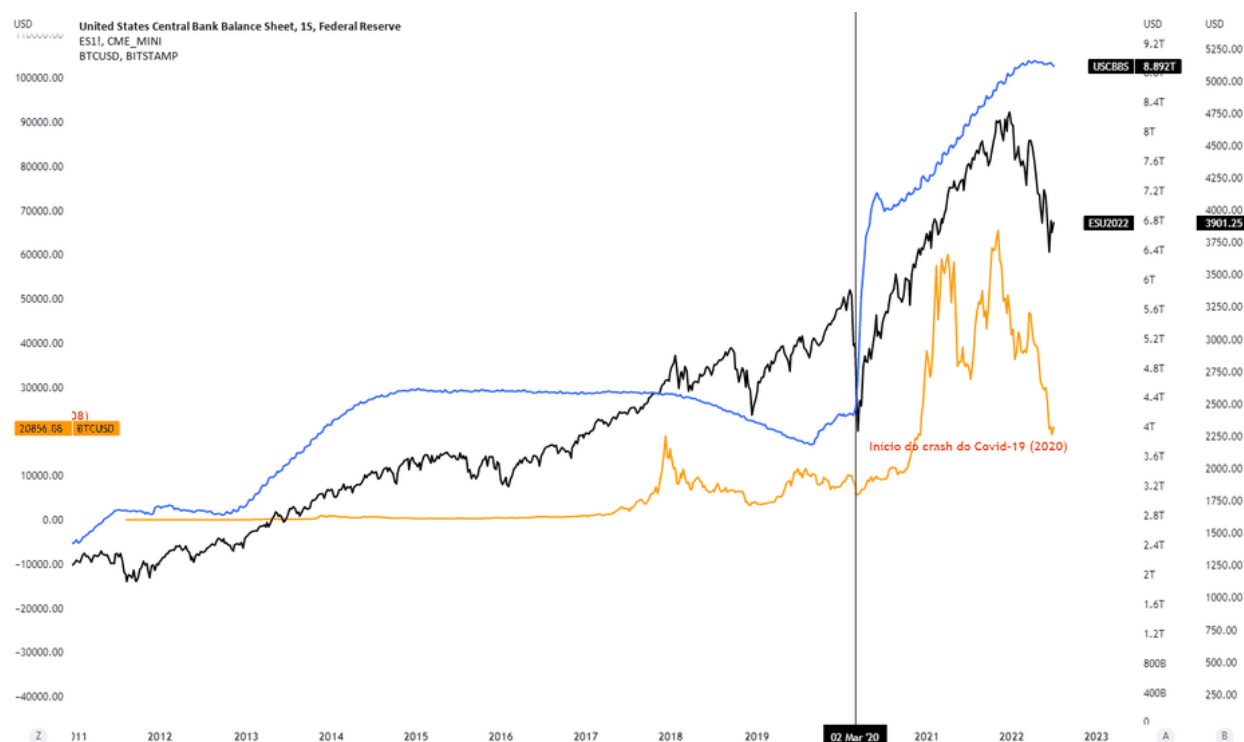
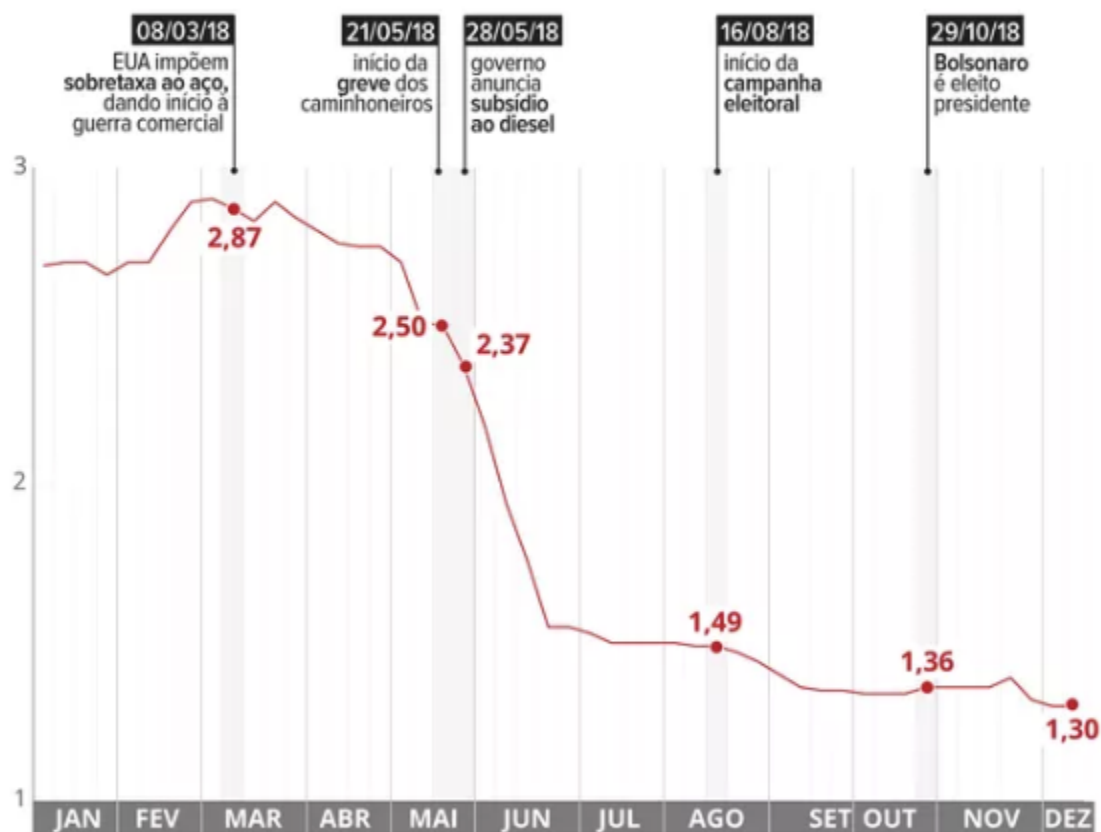


Fig 2: <https://br.tradingview.com/markets/world-economy/>

## Crescimento

### Expectativa para o desempenho do PIB em 2018

Dados em %



Fonte: Banco Central



Infográfico atualizado em: 17/12/2018

Fig 3: <https://g1.globo.com/retrospectiva/2018/noticia/2018/12/21/retrospectiva-2018-a-economia-brasileira-em-6-graficos.ghtml>

**FIM!**