

Navegando na linha do tempo: Uma análise de Séries Temporais.

UMA APRESENTAÇÃO POR JESSICA BRITO



Quem sou eu?

BACHAREL EM
CIÊNCIA DA
COMPUTAÇÃO



PÓS-GRADUADA EM
DATA ANALYTICS



CIENTISTA DE DADOS
JÚNIOR NO
SIDIA AMAZON LAB



FOTÓGRAFA
(AMADORA)

O objetivo é apresentar um
pouco mais sobre a

Área de Dados



A verdade é que vivemos na era da informação: dados são gerados o tempo todo.

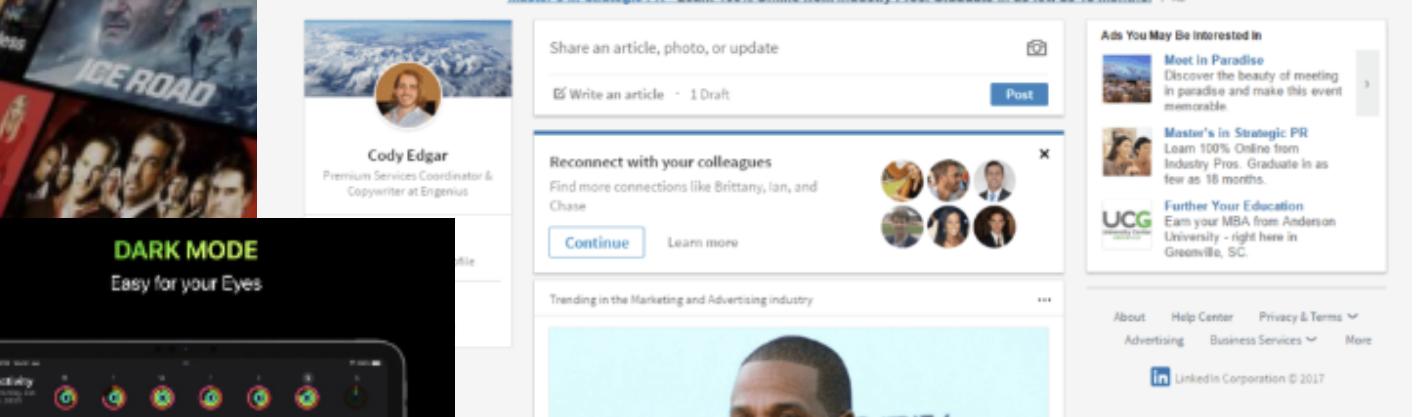
SEJA POR REDES SOCIAIS,
SENSORES DE DISPOSITIVOS, SISTEMAS
FINANCEIROS OU PLATAFORMAS DE STREAMING.



Hyperparameter Tuning and Model Selection

So evaluating a model is simple enough: just use a test set. Now suppose you are hesitating between two models (say a linear model and a polynomial model): how can you decide? One option is to train both and compare how well they generalize using

you want to apply some you choose the value of 0 different models using d the best hyperparamet error, say just 5% error.



del evaluations take. Conversely, it will be much smaller. A del will be trained on a set of dels trained on the training set. In order to participate in cross-validation, each model is evaluated on a validation set, after it is trained on the rest of the data. By averaging out all the evaluations of a model, we get a much more accurate measure of its performance. However, there is a drawback: the training time is multiplied by the number of validation sets.

Mas os dados por si só não são valiosos.

O QUE REALMENTE IMPORTA É
CONSEGUIMOS EXTRAIR DELES
PARA GERAR **INSIGHTS**, PREVISÕES
E TOMADAS DE DECISÃO.



Hyperparameter Tuning and Model Selection

So evaluating a model is simple enough: just use a test set. Now suppose you are hesitating between two models (say a linear model and a polynomial model): how can you decide? One option is to train both and compare how well they generalize using

you want to apply some
you choose the value of
0 different models using
d the best hyperparam-
n error, say just 5% error.

del evaluations to take. Conversely, it will be much smaller. A del will be trained on a validation set, while the other dels trained on the rest of the data. In order to participate in cross-validation, each model is evaluated on the validation set, after it is trained on the rest of the data. By averaging out all the evaluations of a model, we get a much more accurate measure of its performance. However, there is a drawback: the training time is multiplied by the number of validation sets.

Três principais áreas

(QUE PODEM SER DIVIDIDAS EM OUTRAS PEQUENAS ÁREAS)



Análise de Dados

TRANSFORMAR DADOS
BRUTOS EM
INFORMAÇÕES ÚTEIS
PARA O NEGÓCIO



Ciência de Dados

APLICAR ESTATÍSTICA E
MACHINE LEARNING
PARA ENCONTRAR
PADRÕES E PREVER
TENDÊNCIAS



Engenharia de Dados

CRIAR E MANTER A
INFRAESTRUTURA PARA
ARMAZENAR E
PROCESSAR GRANDES
VOLUMES DE DADOS

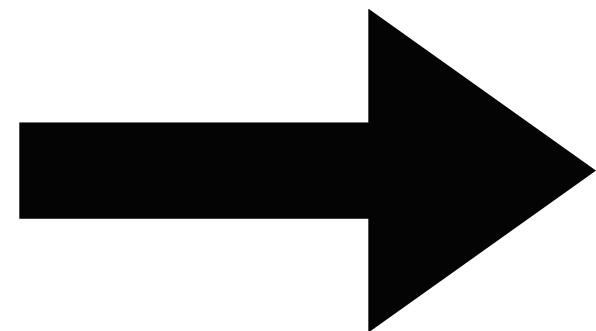
Tópico dentro de Ciência de Dados



Ciência de Dados

APLICAR ESTATÍSTICA E
MACHINE LEARNING
PARA ENCONTRAR
PADRÕES E PREVER
TENDÊNCIAS

Tópico dentro de Ciência de Dados



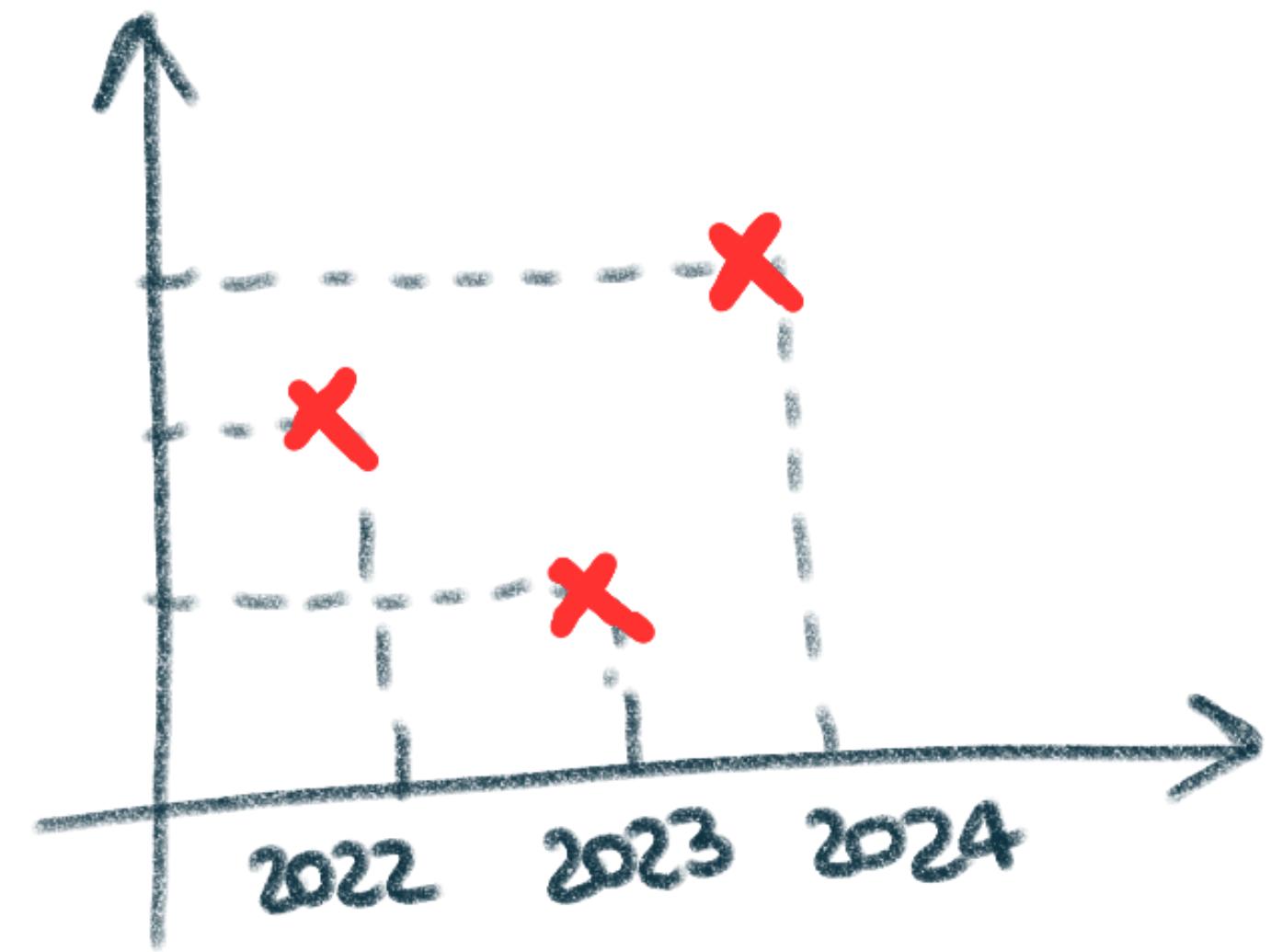
Ciência de Dados

APLICAR ESTATÍSTICA E
MACHINE LEARNING
PARA ENCONTRAR
PADRÕES E PREVER
TENDÊNCIAS

Séries Temporais



Séries Temporais
é um conjunto de
dados ordenados
no tempo.



o que seriam esses dados?

Cientistas sociais
acompanhando taxas de
natalidade e mortalidade
durante uma temporada.

Epidemologista
interessado em observar
o número de casos de
gripe num determinado
período.

Um médico avaliando a
medição da pressão
arterial de um paciente ao
longo do tempo.

A análise de séries temporais nada mais é que navegar na linha do tempo e tirar informações úteis (estatísticas) de dados organizados em ordem cronológica.



**E não só nos perguntar “O que esses dados me dizem?”
mas também “O passado pode ajudar a prever o futuro?”**



**Mas o que de fato seria
essa “análise”?**

**Uma visão
clássica de
análise:**

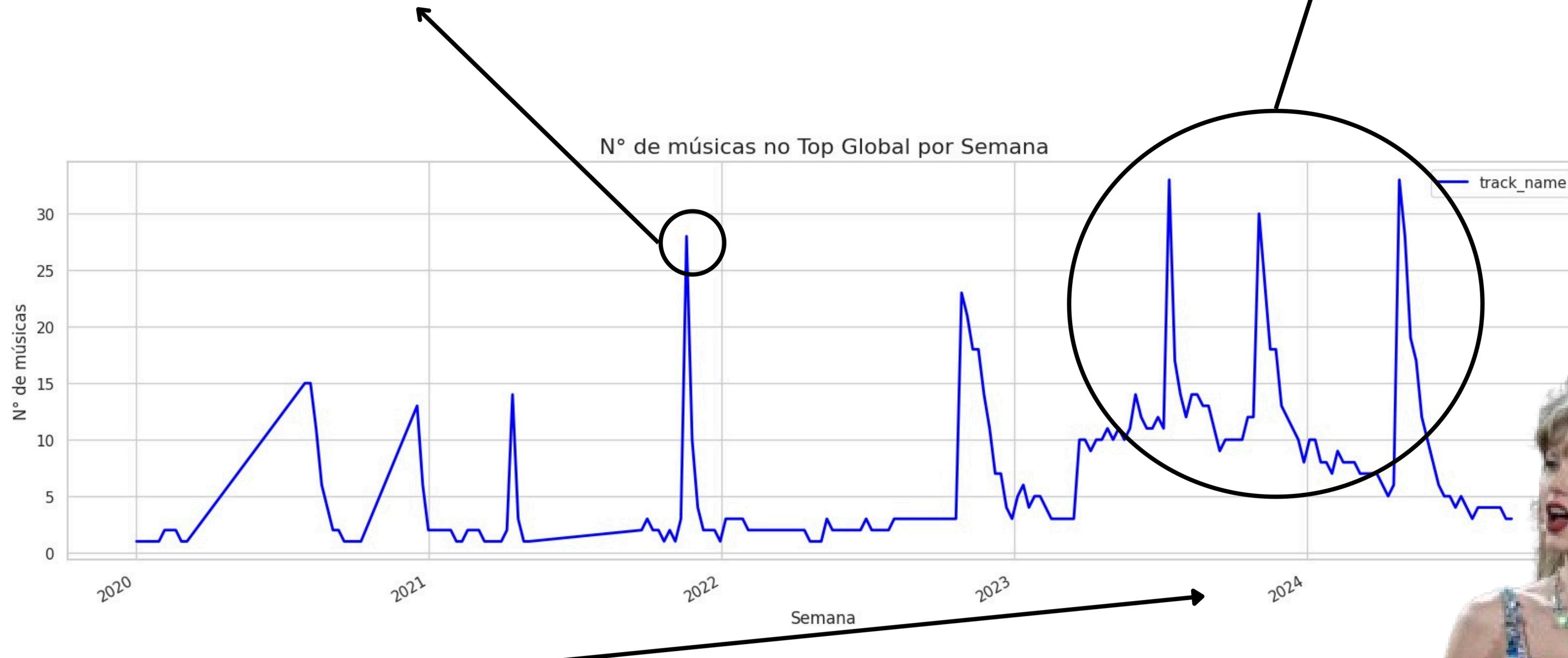
**Uma Série
Temporal é uma
composição de
Tendência,
Sazonalidade e
Ruído.**



Como assim “composição”?

Esse ciclo se repete nos próximos meses?

Qual a explicação para esses picos? Pontos fora da curva?



a tendência é o número de músicas aumentar?

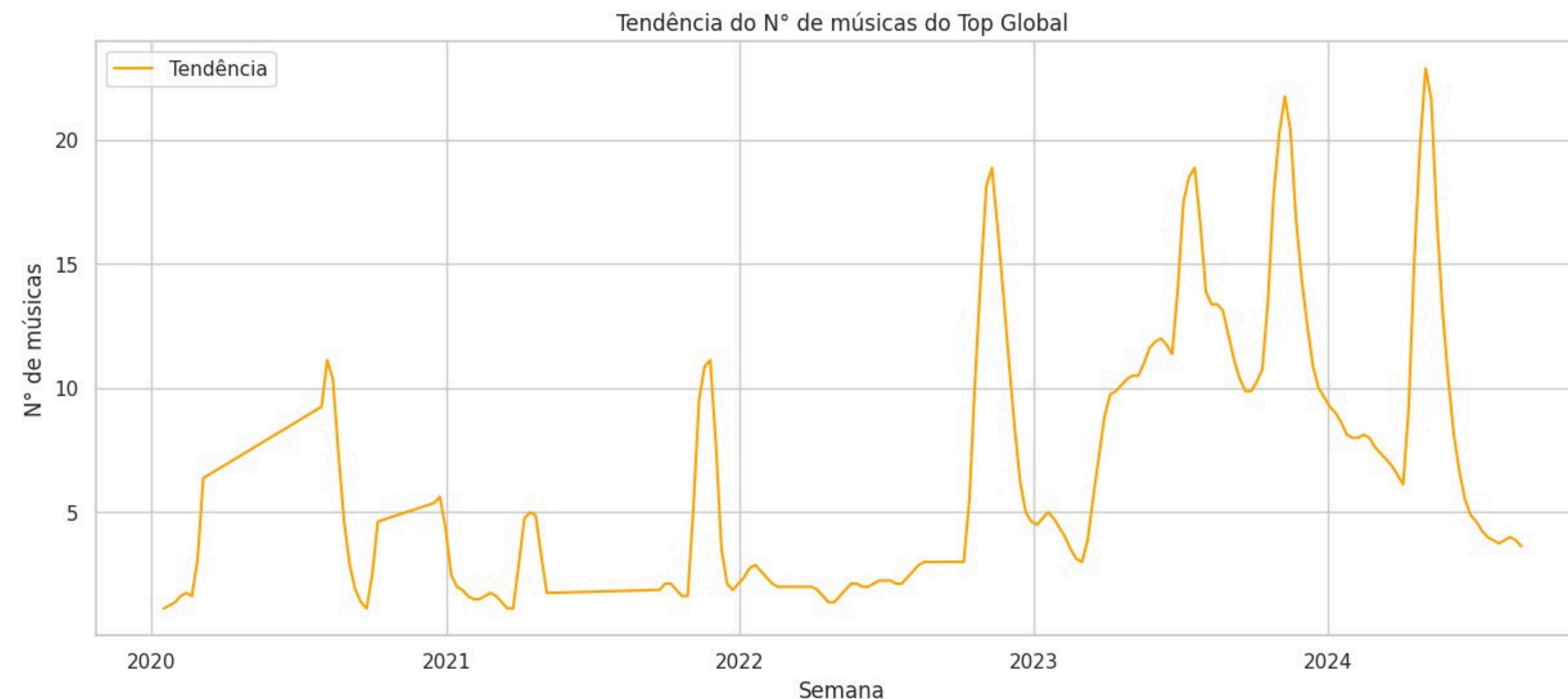


**Para um mesmo conjunto
de dados observados,
podemos tirar diferentes
informações que nos
ajudem a prever dados
futuros.**

**O que seriam esses
componentes?**

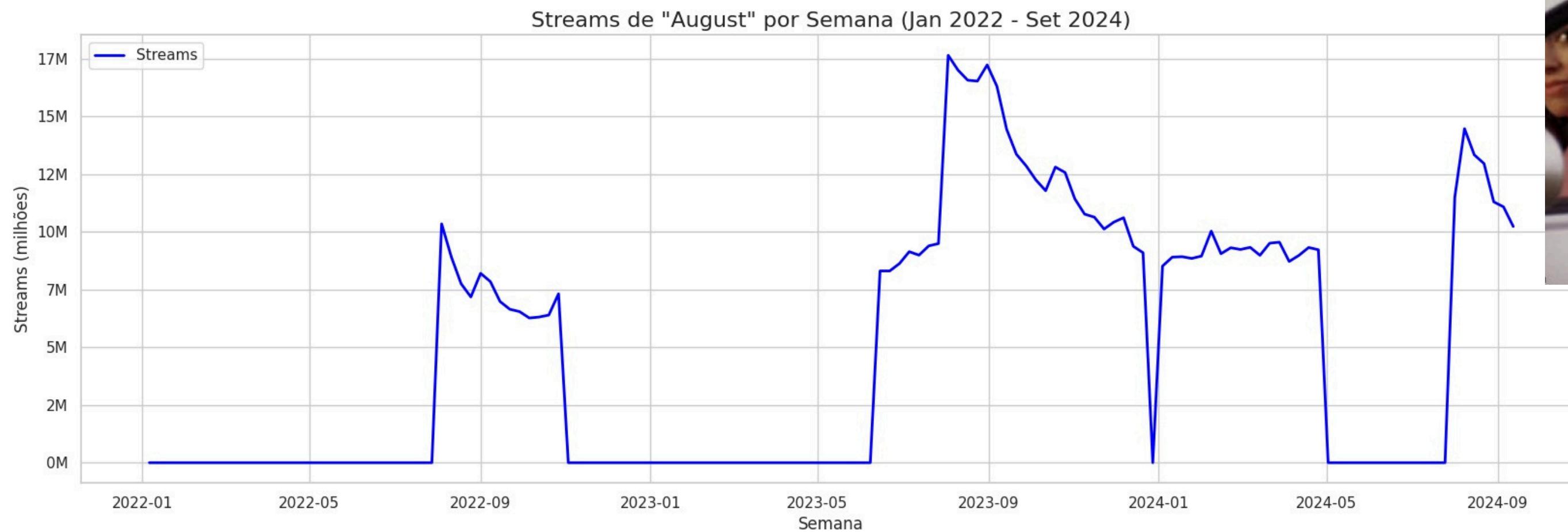
1. Tendência

Direção geral dos dados ao longo do tempo.



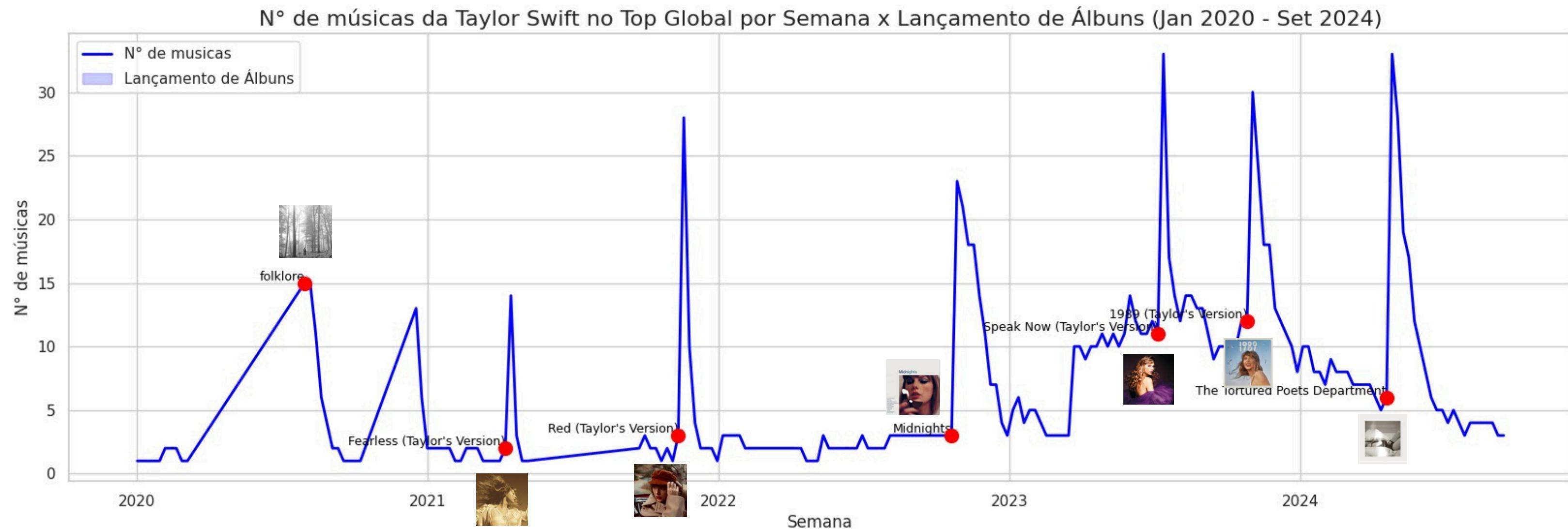
2. Sazonalidade

Variação de dados que sempre ocorre (se repete) em um período específico conhecido.⁵



3. Ruído

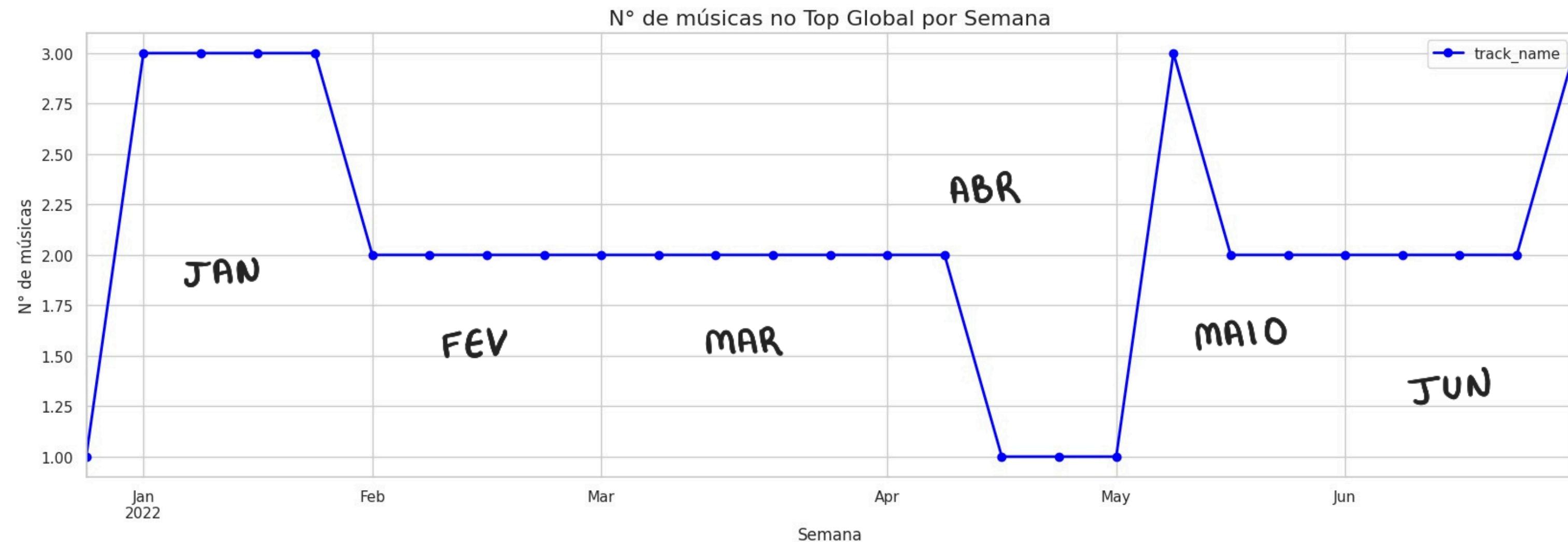
Variação aleatória dos dados que *não* pode ser explicada por nenhum dos outros componentes.



Quando falamos de prever o futuro, é muito mais fácil fazer isso se nossos dados do passado se comportam de forma “constante”.

Vamos observar o período de Janeiro a Junho de 2022:

(um dos raros momentos que essa mulher não estava lançando album)



Duração: 6 meses

**Essa é a média de
músicas no Top
Global por mês:**

Jan	3
Fev	2
Mar	2
Abril	1.5
Maio	2
Jun	2.2

**Neste caso, não estamos errados em olhar para essa
tabela e assumir que - para esse período de 6 meses - a
média é de certa forma constante.**

**Estacionário significa algo
que estacionou ou se
mantém no mesmo lugar
imóvel, parado.**

**Em Séries Temporais, significa que
as propriedades estatísticas
- como média e variância -
se mantém constantes
ao longo do tempo.**

E no que isso implica?

Na maioria dos dados do mundo real, as propriedades estatísticas variam demais com o tempo.

E essas variações podem prejudicar a precisão preditiva dos modelos.

Conseguimos ver a diferença se incluirmos na análise o período no qual houve o lançamento de algum álbum.



Neste cenário, temos uma variação perceptível entre os meses de outubro e novembro.

Jun	2
Ago	3
Set	3
Out	8
Nov	17.75
Dez	6.4

Não precisamos (nem devemos) recorrer a calcular média, variação manualmente para descobrir se uma série é estacionária ou não.

Para isso podemos usar testes estatísticos, dentre eles um dos mais conhecidos:

Augmented Dickey-Fuller Test (ADF Test ou ADFuller)

Augmented Dickey-Fuller Test

(ADF Test ou ADFuller)

- **Hipótese nula (H0) :** Série não é estacionária.
- **Hipótese Alternativa (H1) :** Série é estacionária.



Antes de realizar o teste, você estabelece um nível de significância (probabilidade de cometer um erro) que será usado como medida de comparação para determinamos qual das hipóteses é válida.

Exemplos: 1%, 5% ou 10%

Augmented Dickey-Fuller Test

(ADF Test ou ADFuller)

O teste retorna um valor-p.

Se esse valor for menor ou igual ao nível de significância escolhido (por exemplo, 0.05), rejeitamos a H₀. A série é estacionária.

Mas se for maior, não podemos rejeitar a H₀. Ou seja, a série não é estacionária.



“E a parte mais divertida?”

**Após realizar a análise dos nossos dados,
vamos falar de modelos de previsão.**

Vamos começar com uma previsão simples...

**O último valor registrado é o
valor previsto!**

**Se o último valor registrado foi da
semana 01/01/2024 e eu quero prever
os valores das próximas 8 semanas, os
valores previstos vão ser o valor
registrado da semana 01/01/2024.**

Esse é o modelo **Naive Approach ou modelo “ingênuo”. Ele parte da premissa que “o futuro repetirá o passado” servindo como uma baseline inicial.**

Baselines são métricas comparativas.

Uma particularidade para essa biblioteca - e para algumas outras que tratam de Série Temporais: Ao passar seus dados para construção do modelo, eles precisam ser compostos de três colunas com nomenclaturas específicas.

ds	unique_id	y
Dados cronológicos	Identificador para um ou mais tipos de dado sendo observados	Dados observados/ coletados

**Para mensurar quão bem seu modelo
tem predito novos valores:**

MAPE



MAPE

Erro Absoluto Percentual Médio

- **Começa calculando o erro absoluto percentual para cada ponto de dado**
- **Soma todos os erros e divide pelo número total de amostras (média)**
- **Pode multiplicar por 100 para ficar mais fácil de explicar!**

Como interpretar?

↓ MAPE ↑ MAPE



**Seu modelo está
prevendo bem!**

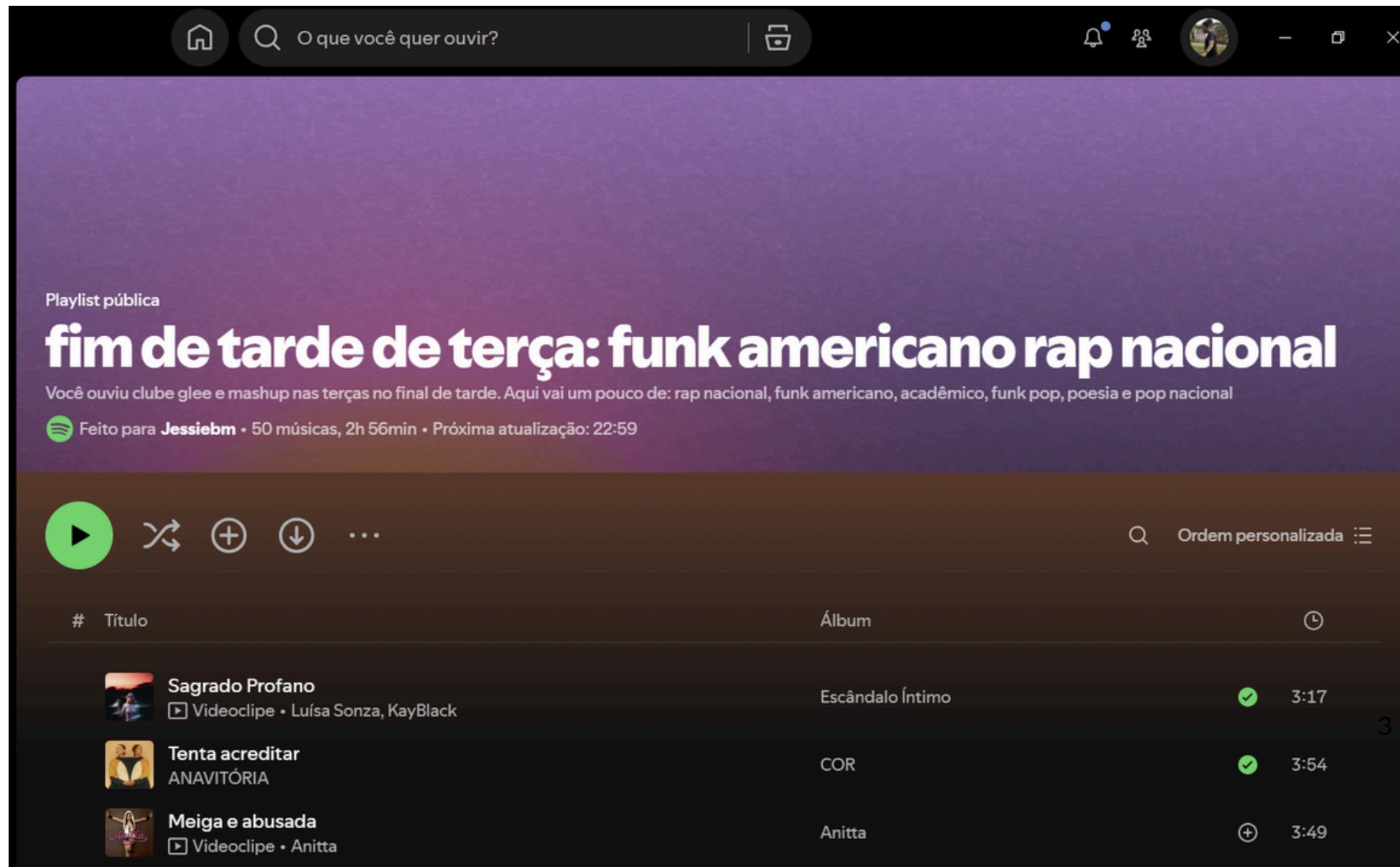


**Seu modelo não está
prevendo bem!**

É pertinente dizer que os resultados podem variar de acordo com o os valores que escolhemos submeter para treino e o quanto escolhemos prever do futuro.

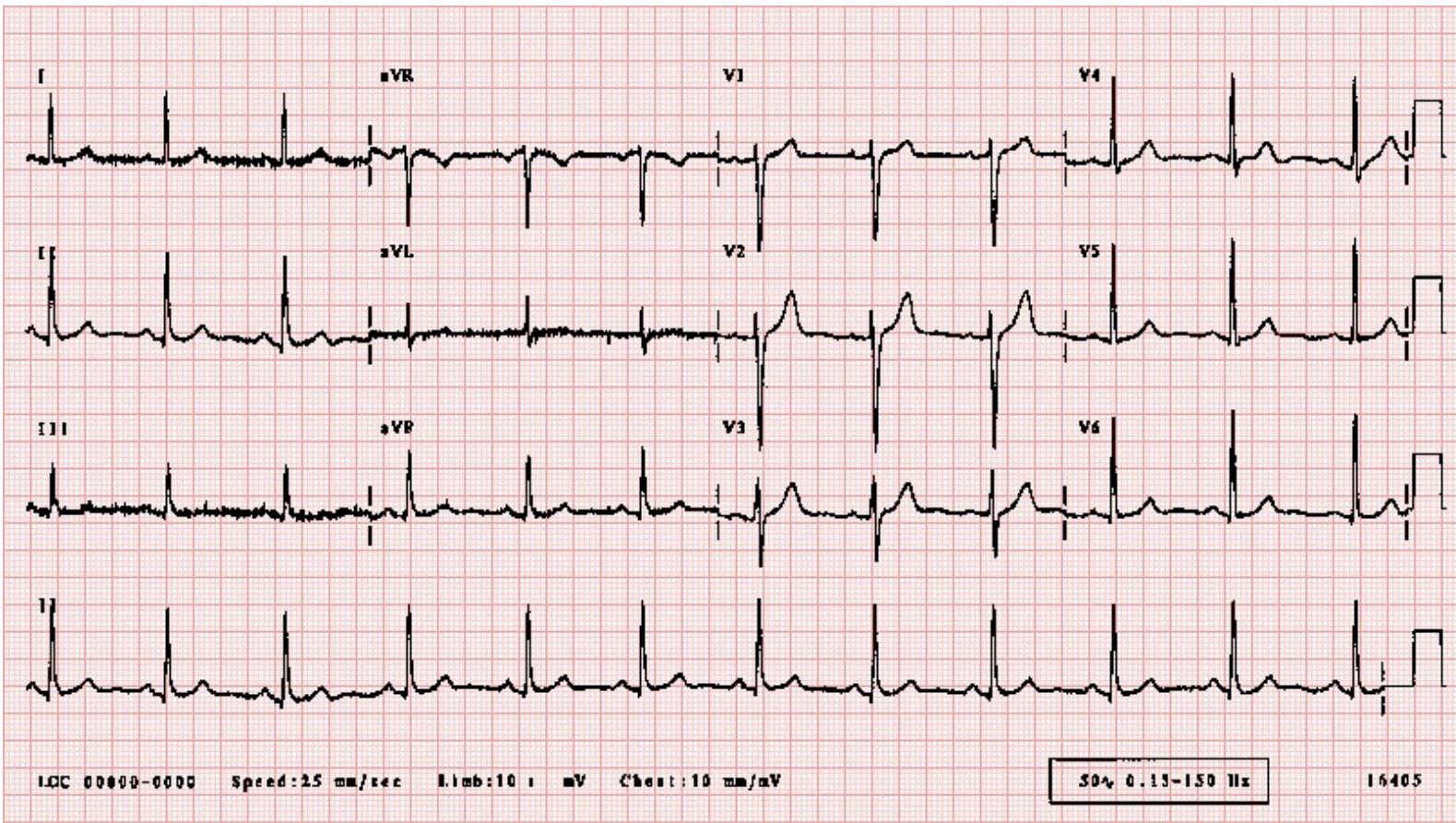
Para este caso, tentamos prever os próximos 3 meses. Mas e se tentássemos prever apenas o dia seguinte? Ou prever os próximos 6 meses?

MÚSICA



Uma nova funcionalidade oferecida pelo Spotify é criar playlists de música de acordo com o dia da semana e o período do dia.

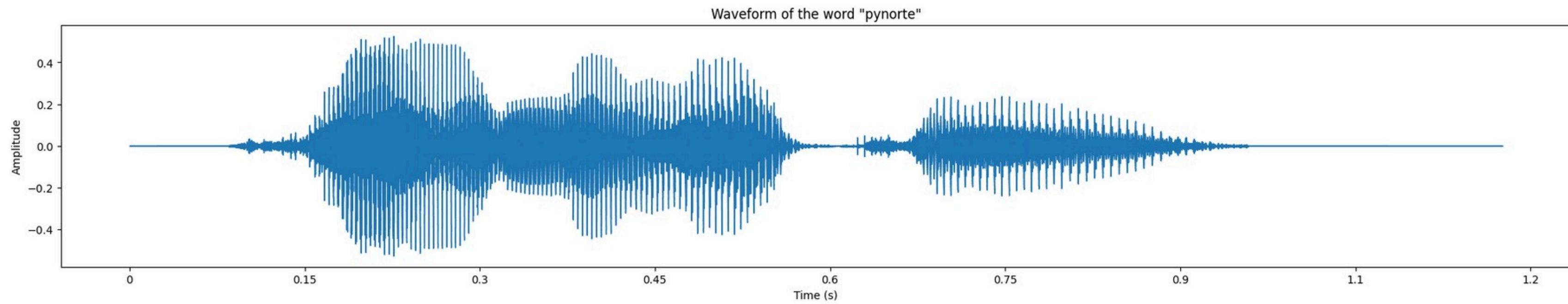
MEDICINA



A análise de séries temporais conseguiu seu espaço na medicina quando os primeiros eletrocardiogramas funcionais (ECGs), que conseguiam diagnosticar condições cardíacas registrando os sinais elétricos que passam pelo coração, foram inventados em 1901

O ECG continua sendo uma área atuante de pesquisa, cuja finalidade é muito prática, como estimar o risco de uma crise cardíaca súbita ou uma convulsão.

PROCESSAMENTO DE LINGUAGEM



Pode ser realizada uma decomposição de dados para produzir uma “assinatura” que pode ser comparada com assinaturas de diferentes sílabas em uma biblioteca para encontrar uma correspondência.

Previsão de Séries Temporais: Abordagem Além dos Números



PT-BR



Iniciante

Guia rápido para iniciantes em séries temporais



PT-BR



Iniciante



EN



Intermediário

An introduction to non stationary time series in python

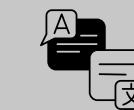


EN



Iniciante

Como prever séries temporais univariadas usando python



PT-BR



Iniciante

Forecasting principles and practice



EN



Intermediário

Obrigada!



/in/jessica-brito-moura/



jessibmoura



@jcabrito

