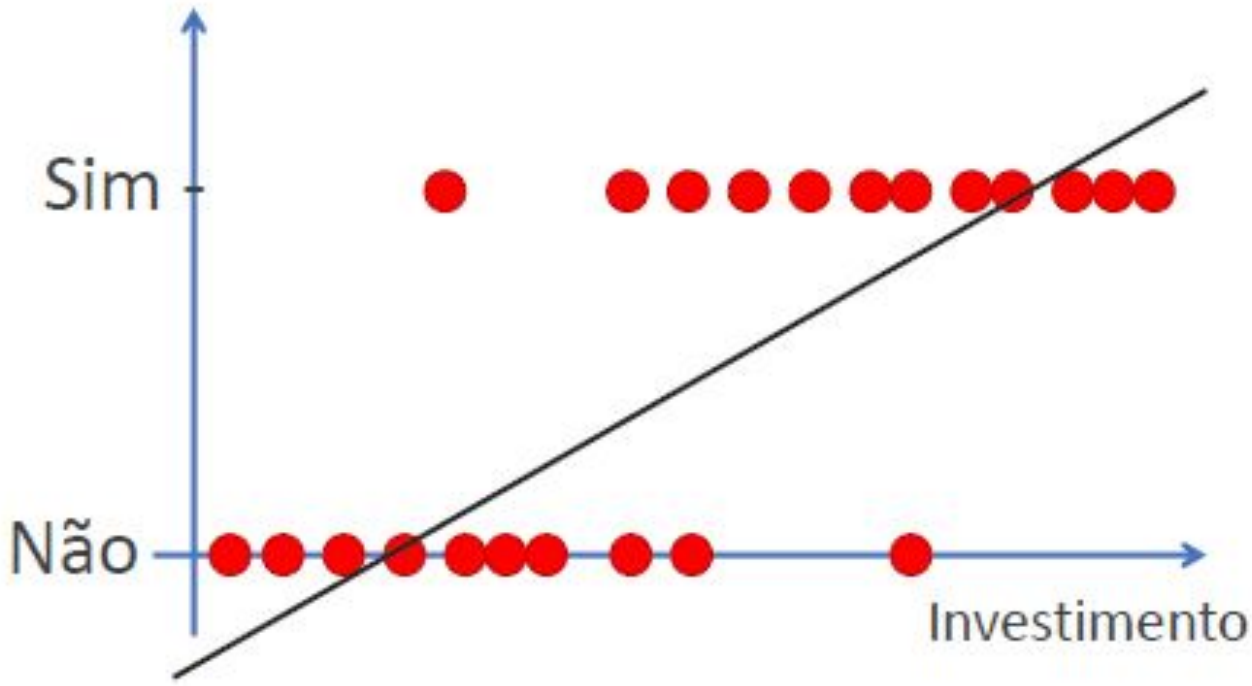
DATA SCIENCE

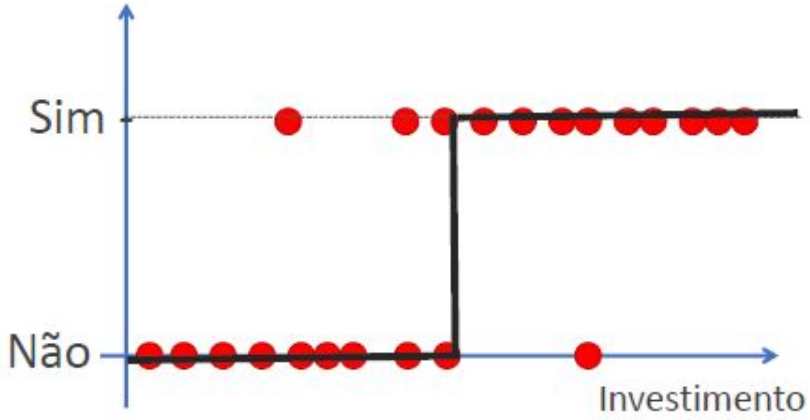
AULA 11 – SISTEMA DE RECOMENDAÇÃO II

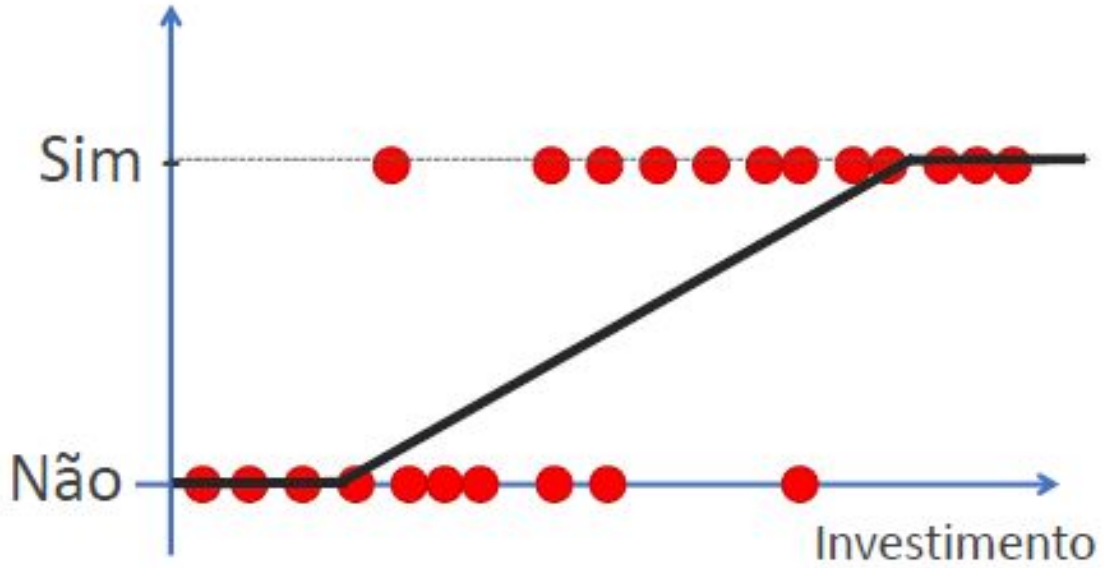
PROFª. ANA CAROLINA B. ALBERTON

REGRESSÃO LOGÍSTICA



REGRESSÃO LOGÍSTICA





REGRESSÃO LOGÍSTICA

Semelhante a regressão Linear, porém a variável de resposta é binária: sucesso ou fracasso

1: sucesso

0: fracasso

O sucesso ou fracasso é representado através de probabilidade

Também pode ser simples ou múltipla

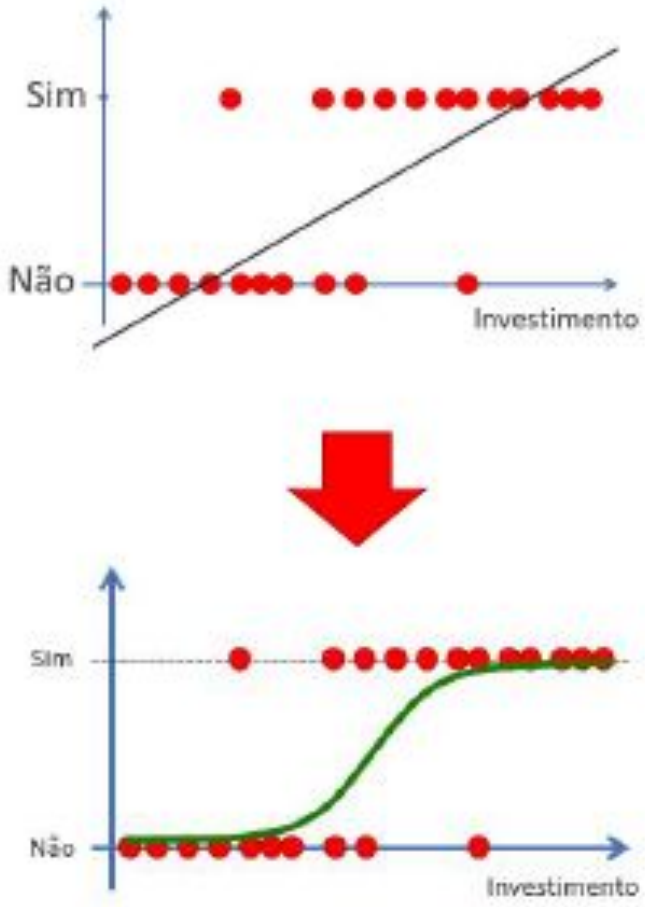
REGRESSÃO LOGÍSTICA

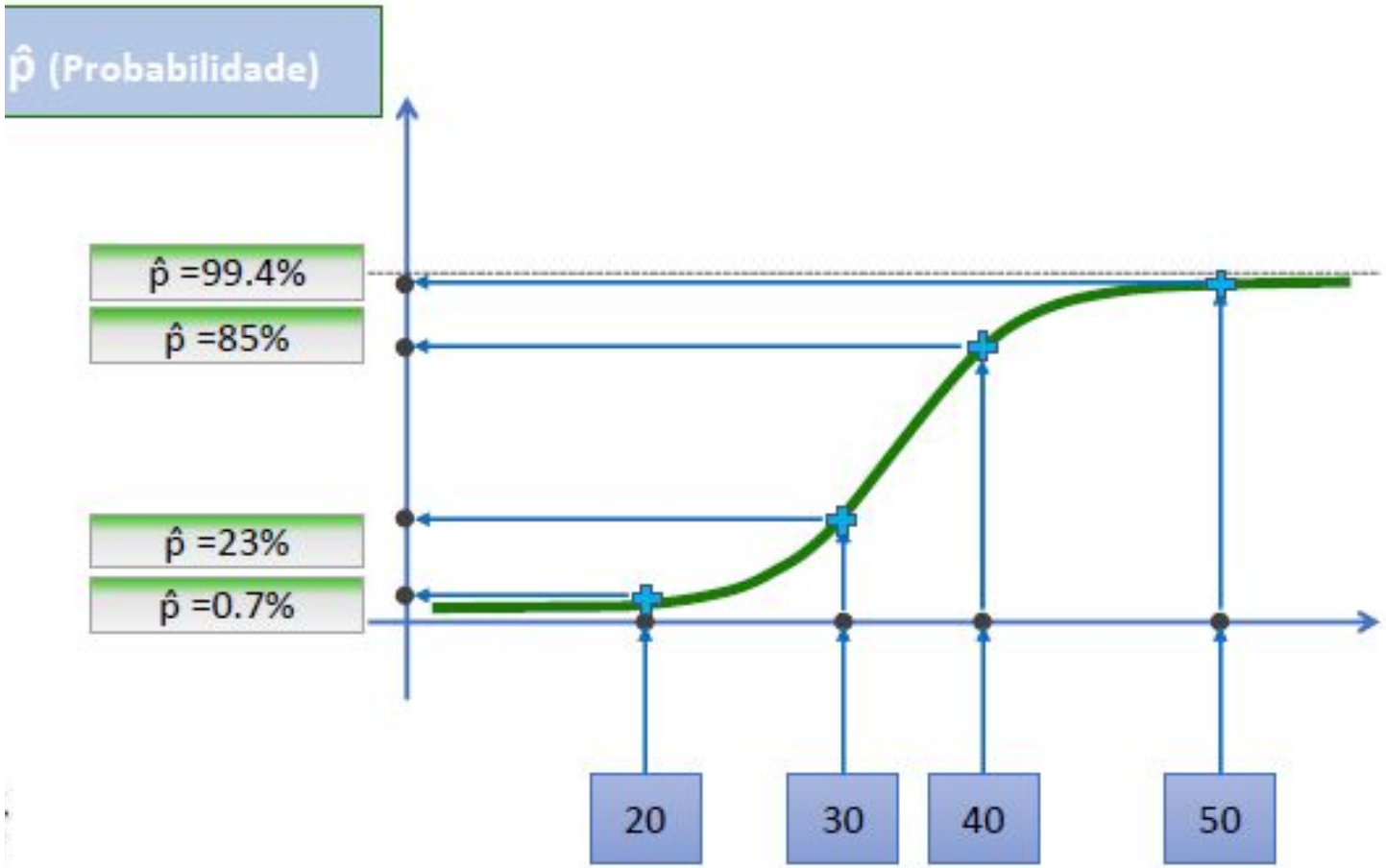
$$y = b_0 + b_1 * x$$

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$

Melhor representada pela função Sigmóide




p̂ (Probabilidade)
p̂ =99.4%
p̂ =85%
p̂ =23%
p̂ =0.7%
20
30
40
50

SÉRIES TEMPORAIS

- Até o presente momento vimos como identificar fórmulas para descrever e prever fenômenos que julgamos serem contínuos e imutáveis, através das regressões lineares (simples e múltiplas) e não lineares;

- Isso pode funcionar para várias outras questões e de fato é primeiro ponto quando falamos em predições;
- Contudo, nessa etapa da disciplina iremos focar em predições com distribuições de séries temporais;

SÉRIES TEMPORAIS

- Uma série temporal é o nome que se dá à uma distribuição de dados, os dados são distribuídos em função de um período temporal;

- Portanto, quando queremos entender ou prever comportamentos cujo o fenômeno esteja diretamente relacionado à um fator temporal, trabalharemos com séries temporais;

- Fatores temporais podem ser: Hora, Dia, Mês, Ano. Ou seja, qualquer "momento" temporal, onde os dados podem ser registrados e posto em uma ordem.

SÉRIES TEMPORAIS

- Grandes decisões envolvendo Data Science podem ser modeladas com séries temporais. Aqui vão alguns exemplos:

  - "Uma subestação de energia quer entender em qual momento do dia deve-se aumentar a capacidade de geração de energia para seus usuários";

  - "Uma loja quer entender qual o estoque ideal que ela deve se abastecer para o verão";

  - "Uma pizzaria que saber com antecedência a quantidade de massa que deve ser pré-pronta de acordo com o dia da semana".

SÉRIES TEMPORAIS

- Notem que em cada um dos casos que eu dei como exemplo, foi utilizado um fator temporal diferente e uma situação diferente, porém todas elas possuem de um certo modo o mesmo padrão;

- Temos uma variável que quantifica o tempo, ou melhor que determina o momento que algo acontecer, e temos outra variável que quantifica o fenômeno (seja ele qual for);

- Portanto, mesmo que tenhamos aqui uma data, no fim, o que temos em resumo seriam duas variáveis com valores numéricos.

SÉRIES TEMPORAIS

- As series temporárias podem ser:
  - Estacionárias – flutuam em torno de uma mesma média e variância;
  - Não estacionárias.
  - Estocásticas – formula + fator aleatório que não pode ser explicado;
  - Determinísticas – Explicadas através de uma formula/função.

SÉRIES TEMPORAIS - TIPOS

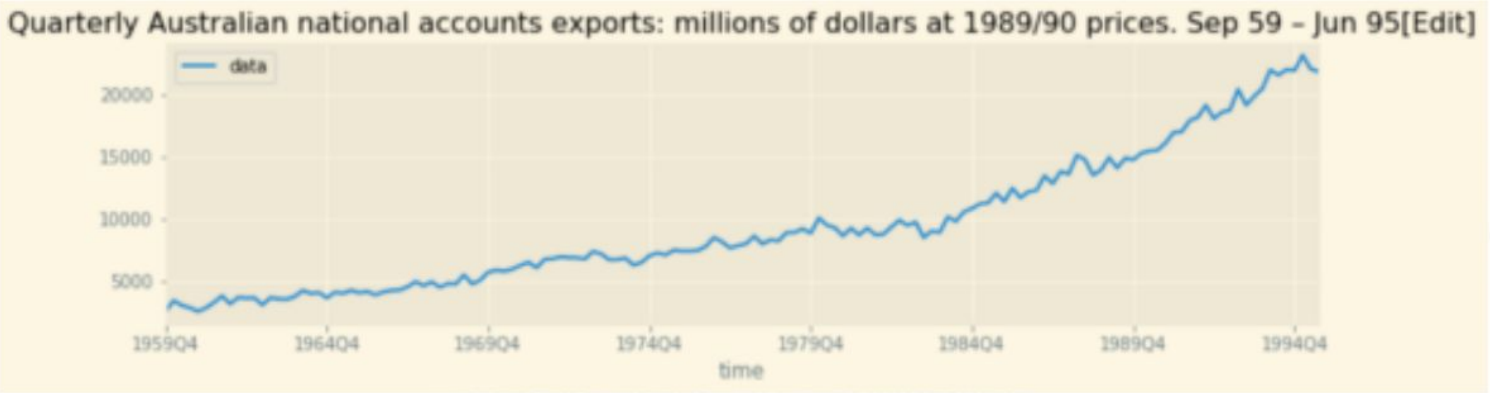
- Em toda distribuição de série temporal ao analisarmos ela graficamente é possível extrair delas algumas características;

- Essas características nos ajudam a tipificar as séries temporais em 4 tipos principais de séries, são elas:

  - Tendência;
  - Sazonalidade;
  - Randômica.

TENDÊNCIA

- Uma série que apresenta uma tendência irá se parecer mais ou menos com o seguinte gráfico:

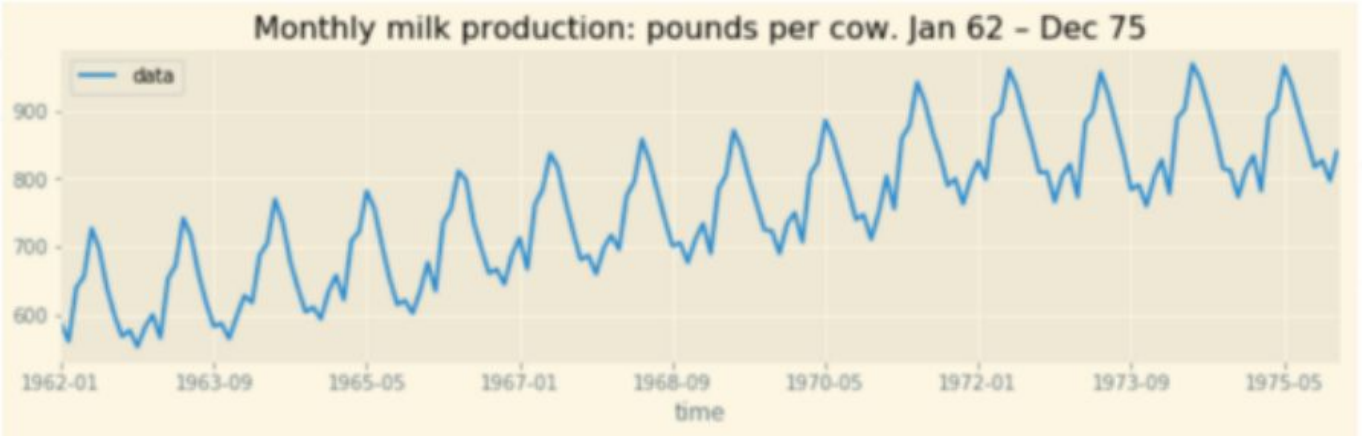


- As séries temporais com tendência, como o nome já diz apresentam em um determinado momento uma evolução ou decréscimo constantes;

- Isso inclusive determina como devemos interpretar o seu futuro comportamento, ou melhor como os modelos irão interpretá-lo.

SAZONALIDADE

- Uma série que apresenta uma sazonalidade irá se parecer mais ou menos com o seguinte gráfico:



- Quando falamos sobre sazonalidade, o que podemos observar é que há um padrão que ocorre durante um período de tempo e que se reinicia ao finalizar o ciclo.

SAZONALIDADE

- Peguemos por exemplo viagens de avião para um país que possua as quatro estações bem definidas:

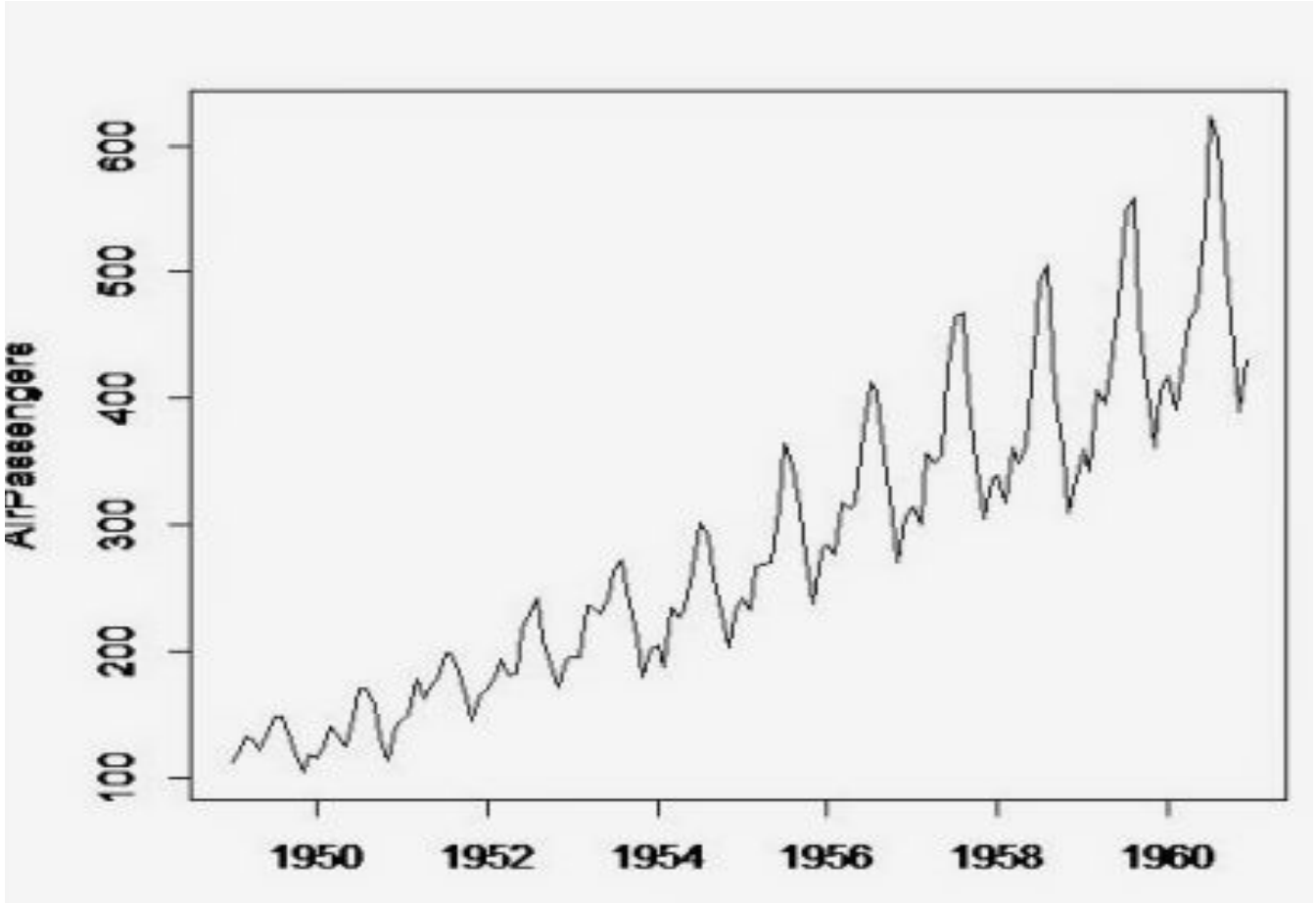
- Quando é alta-temporada temos um grande número de turistas pelas cidades, o que diminui drasticamente ao chegar o inverno e que recomeça basicamente na mesma proporção ao chegar o verão.

- O que acontece nesses casos, é que os modelos devem aprender em qual fase da sazonalidade ele está, para então identificar qual é o movimento futuro.

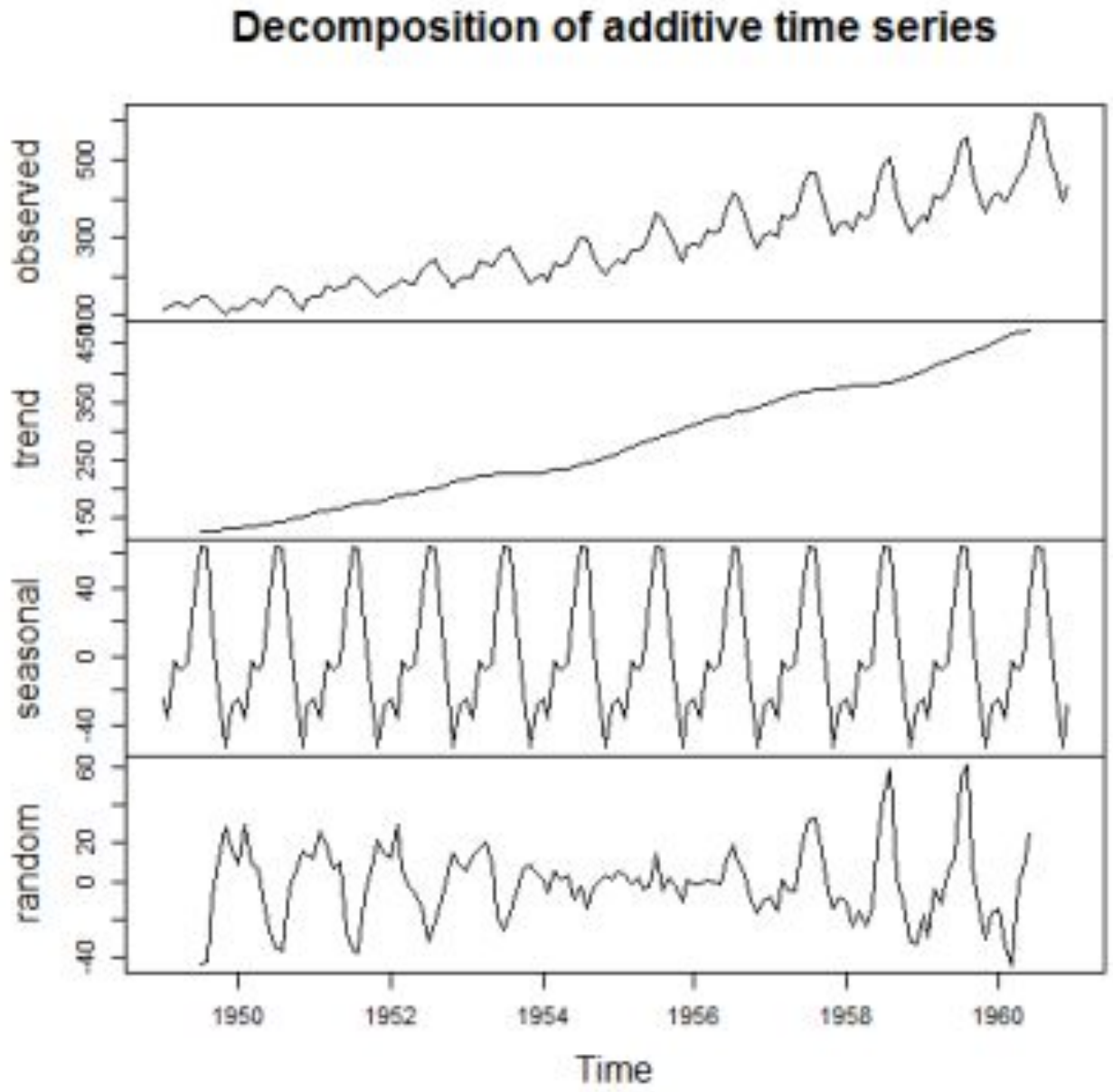
RANDÔMICAS

- Séries randômicas não se parecem com nenhum tipo de padrão, como abaixo:

- Nessa caso dificilmente há uma aprendizagem de modelo, as previsões se tornam cada vez mais incertas, e provavelmente são dependentes de outros fatores além simplesmente da passagem do tempo.


AirPassengers
100
200
300
400
500
600
1950
1952
1954
1956
1958
1960

Ver código em python



SÉRIES TEMPORAIS

- Seguindo portanto nessa ideia, podemos relacionar todo o aprendizado que viemos tendo ao longo do semestre;

  - **Primeiro,** quando vamos trabalhar com séries temporais a primeira decisão é verificar qual fator temporal mais se adequa à serie que se pretende estudar;

  - Não podemos deixar de envolver o contexto nessa jogada, o que é bem óbvio, sem o conhecimento prévio do contexto da série nada se pode fazer;

SÉRIES TEMPORAIS

- **Segundo,** nós temos que verificar a normalidade de nossa distribuição, analisar fatores que poder ter ocorrido apenas uma única vez e, que podem distorcer totalmente a nossa análise;

- Então também é justo dizer que a análise de outliers vai lhes auxiliar muito quando formos trabalhar com as séries temporais;

- **E por fim**, uma lição que vimos com a regressão linear é que, quando apenas uma única variável não é suficiente para entendermos um fenômeno, existe sempre a possibilidade de incluirmos novas variáveis afim de melhorar nossa análise

MODELOS

- Vamos então ver alguns de modelos que tratam séries temporais.
  - Como modelos mais simples de serem implementados temos:
    - SES (Simple Exponential Smoothing):
    - Holt – Linear:
    - Holt – Winter:

SES – SIMPLE EXPONENTIAL SMOOTHING

- A suavização exponencial simples é um dos modelos mais simples a serem utilizados para séries temporais;
- O seu uso é indicado quando temos poucos dados e, além disso, esses dados forem irregulares, sem apresentar uma aparente tendência ou sazonalidade;
- Para esse modelo uma média ponderada é calculada através das observações anteriores que vai sendo balanceada ao longo da distribuição;
- Nesse caso, a previsão do próximo valor será basicamente o valor dessa média ponderada.

HOLT LINEAR

- O modelo Holt Linear também pode ser considerado um processo de suavização da distribuição porém esse modelo tem o incremento de um componente de tendência;
- Portanto o seu uso é indicado para tratar **distribuições que apresentem tendências**;
- Sendo que para esse caso, além da suavização, o modelo vai aprendendo o "nível" da tendência onde ele se encontra para prever o próximo valor;
- Então, quando a tendência é positiva, o valor previsto será sempre maior que o anterior, ou será menor quando a tendência for inversa.

HOLT WINTER

- O modelo Holt Winter é uma evolução do modelo visto anteriormente;
- Sua principal diferença será que ele foi evoluído para que fosse possível tratar também sazonalidades com esse modelo;
- Portanto temos um modelo para cada necessidade (ou tipo de série) que vimos.

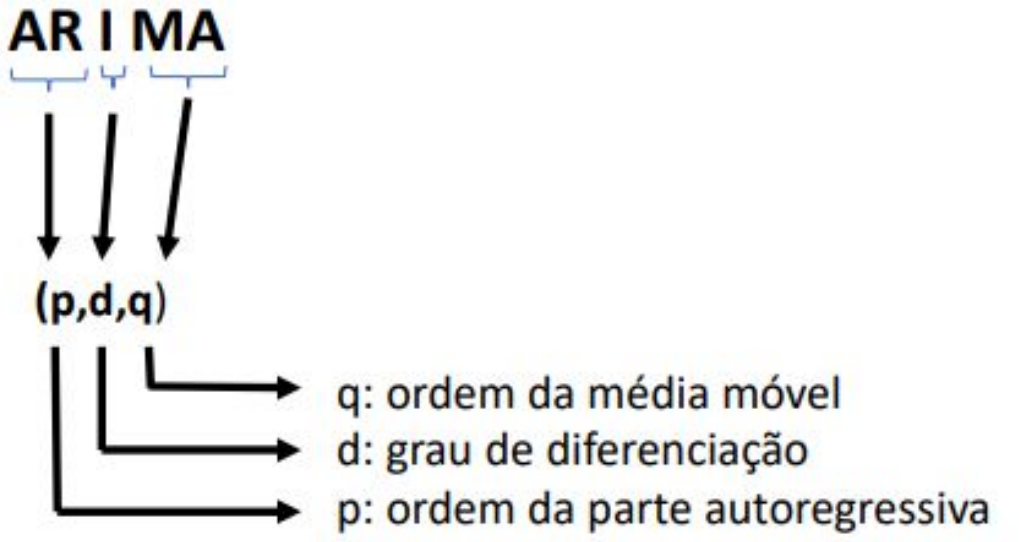
ARIMA

- *AutoRegressive Integrated Moving Average,*
- *Usar em dados estacionários;*
- E é um dos modelos mais utilizados na utilização de previsões por séries temporais;
- Na realidade ARIMA é uma classe de modelos, que compreendem além do ARIMA, modelos como:
  - ARMA;
  - SARIMA;
  - SARIMAX*.

*Seasonal AutoRegressive Integrated Moving Average with eXogenous

ARIMA

Na criação do modelo através do Python, a forma mais básica da sua criação é a passagem numérica de 3 fatores que chamamos de Hiper-Parâmetros.



ARIMA

- A auto regressão (p) seria quantos registros pra trás o modelo vai olhar pra tentar prever o próximo. Seria como utilizar um fato para prever a tendência da série, ou seja, a quantidade de registro que é suficiente pro modelo perceber se está descendo ou subindo.
- A integração (d) nada mais é do que um conceito chamado de diferenciação, portanto, indicar a quantidade de diferenciação pro modelo, é o mesmo de você mesmo aplicar ela antes;
- A média móvel (ma) é dizer para o modelo quantos registros ele precisa olhar pra trás para entender a quantidade média daquele fenômeno;

ARIMA

- Então, temos que decidir quais valores irão ser jogados em p, d e q;
- Esses valores são valores inteiros, que vão de 0 até a quantidade que você achar melhor;
- Só que quanto maiores os valores mais tempo e processamento será consumido, com possibilidade do modelo rodar em minutos, horas, dias e até semanas da forma que for programado;
- Vamos ao código então.

ARIMA

- Como saber qual o melhor modelo?
  - Akaike Information Criteria (AIC e AICc)
  - Baysian Information Criteria (BIC)
- Objetivo é encontrar os menores valores

- Definir os parâmetros p,d e q pode ser extremamente difícil, mesmo para experientes
- Não é um processo linear e nem sempre o modelo intuído é o melhor

- Como fazer? Testar todas as combinações prováveis?

- Usando **Auto.arima()** [em python]
  - Testa diferentes combinações de p,d e r
  - Extremamente flexível
  - Mesmo intuindo um modelo, você pode usá-la para confirmar sua parametrização

- Vamos ao código então.