



Data Science Academy

Big Data Analytics com R e Microsoft Azure Machine Learning Módulo 5



Data Science Academy

www.datascienceacademy.com.br



R Fundamentos

Parte 4



Data Science Academy

Ferramenta	Vantagens	Desvantagens	Utilização	Open Source
R	<ul style="list-style-type: none"> Muitas bibliotecas Excelentes ferramentas para visualização de dados 	<ul style="list-style-type: none"> Curva de aprendizagem maior 	<ul style="list-style-type: none"> Análise Estatística Finanças 	Sim
SAS	<ul style="list-style-type: none"> Suporta grandes conjuntos de dados 	<ul style="list-style-type: none"> Alto custo Linguagem de programação um pouco defasada 	<ul style="list-style-type: none"> Ambiente Corporativo Governos 	Não
Stata	<ul style="list-style-type: none"> Facilita o processo de análise estatística 	<ul style="list-style-type: none"> Armazena os dados em memória 	<ul style="list-style-type: none"> Ambiente Corporativo 	Não
SPSS	<ul style="list-style-type: none"> Primeira versão liberada em 1968 (maturidade) Data e Text Mining 	<ul style="list-style-type: none"> Alto custo Dificuldade de uso 	<ul style="list-style-type: none"> Ambiente Corporativo Governos 	Não
Matlab	<ul style="list-style-type: none"> Manipulação de Matrizes Construção de Plots 	<ul style="list-style-type: none"> Foco maior em matemática computacional e menos em análise de dados 	<ul style="list-style-type: none"> Engenharia 	Não
Scipy, NumPy, Matplotlib	<ul style="list-style-type: none"> Soluções Python, contam com uma grande comunidade de desenvolvimento 	<ul style="list-style-type: none"> Requerem mais tempo de amadurecimento 	<ul style="list-style-type: none"> Engenharia Propósito Geral 	Sim





Comparação entre Pacotes Estatísticos

https://en.wikipedia.org/wiki/Comparison_of_statistical_packages



Data Science Academy

Prós e contras do R vs outros pacotes estatísticos:

- SAS é mais preparado para grandes bases de dados. O R guarda tudo na memória.
- SAS/SPSS tem suporte dedicado e garantia das suas rotinas e ferramentas.
- O R não tem suporte (apenas comunidade que se ajuda) e não possui nenhuma garantia das rotinas feitas por seus colaboradores.
- SAS/SPSS estão no mercado há muito mais tempo que o R.
- SAS é um "canivete suíço". Ele é ferramenta de análise, governança corporativa, de gestão e serviços de TI e muito mais.
- R é gratuito. SAS e SPSS tem alto custo de licenciamento.



Data Science Academy



Web Crawling

x

Web Scraping



Data Science Academy



Web Scraping (Web Harvesting / Web Data Extraction) HTML e CSS



Data Science Academy



robots.txt

Allow: /directory1/myfile.html

Disallow: /directory1/



Data Science Academy

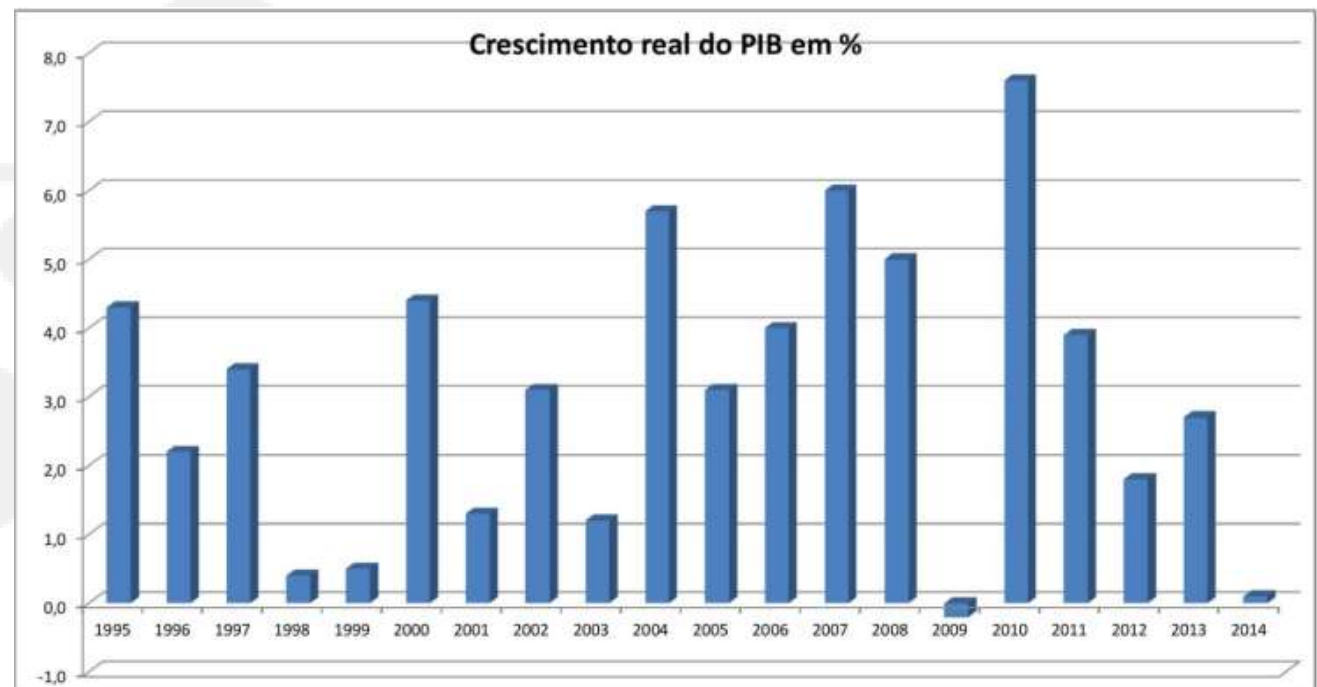
Subsetting

- Os operadores de subsetting
- Os tipos de subsetting
- Diferença de subsetting entre os objetos (vetor, matriz, lista, dataframe)

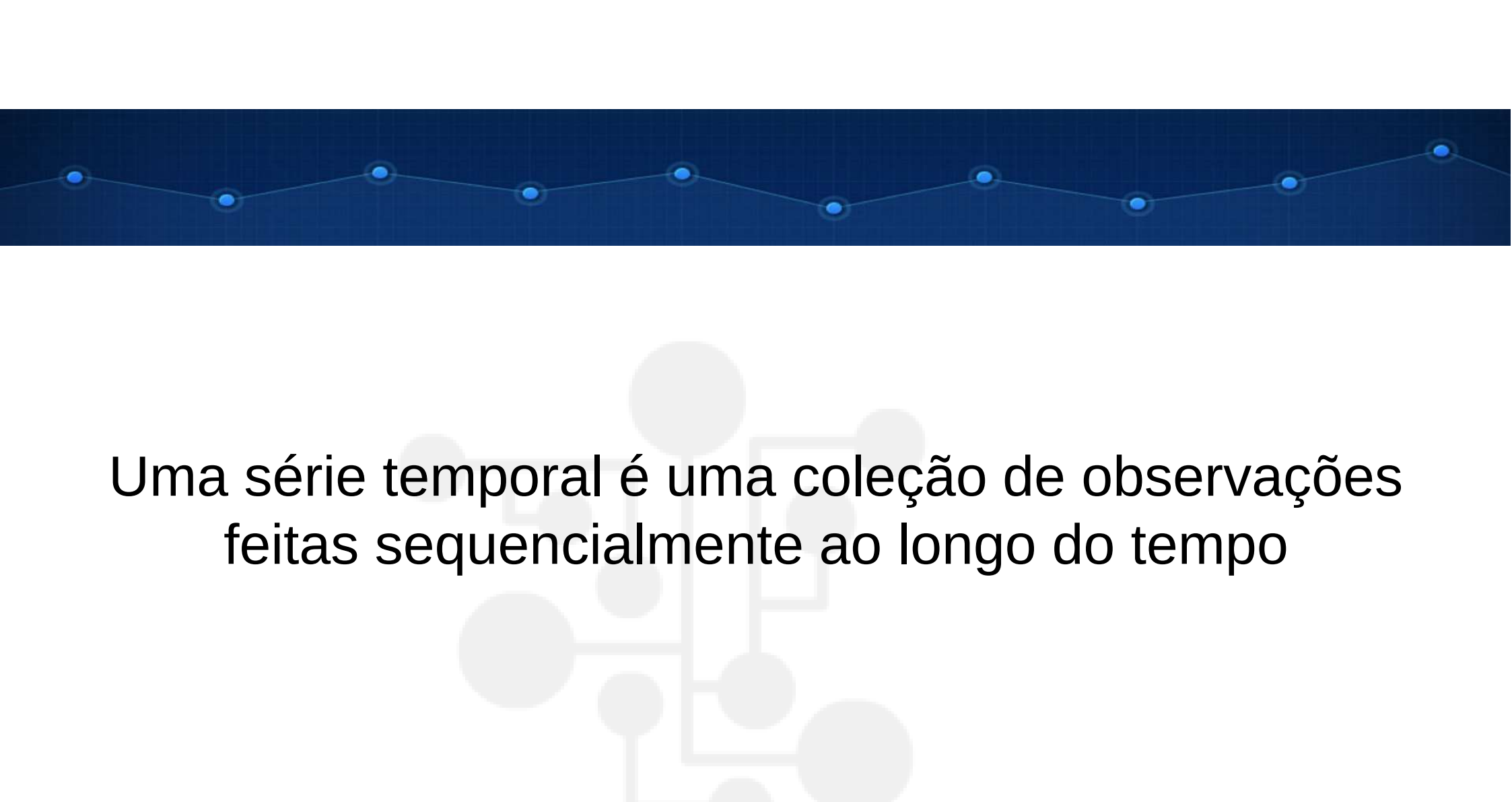


Data Science Academy

Séries Temporais



Data Science Academy



Uma série temporal é uma coleção de observações
feitas sequencialmente ao longo do tempo



Data Science Academy



Ao estudar Séries Temporais, estamos interessados em 2 aspectos:



Análise e
Modelagem

Previsão



Data Science Academy



Análise de Séries Temporais

- Utilizamos dados históricos para descrever a trajetória mais provável da série no futuro.
- Na maioria dos problemas, o passado traz informações relevantes sobre o que irá ocorrer no futuro, pois existe correlação entre as variáveis em diversos instantes.
- É claro que o conhecimento do passado não nos diz exatamente como será o futuro e então sempre existe a incerteza associada às nossas previsões.
- Mas podemos ter uma boa ideia de quais serão os valores mais prováveis no futuro.
- Ou seja, podemos especificar previsões futuras e limites de confiança.



Data Science Academy



Pacotes R para Séries Temporais

Pacote	Descrição
tseries	Séries temporais e análises financeiras
forecast	Modelos de séries temporais
uroot	Testes de raiz unitária
dynlm	Modelos lineares dinâmicos e regressão com séries temporais
dse	Modelos de séries temporais multivariados



Data Science Academy