**Introdução**

Como já esperado, a função de um Cientista de Dados envolve principalmente a exploração e a análise de dados. Embora o resultado final da análise de dados possa ser um relatório ou um modelo de machine learning, os cientistas de dados começam o trabalho com os dados, tendo o Python como a linguagem de programação mais popular usada por eles para trabalhar com os dados.

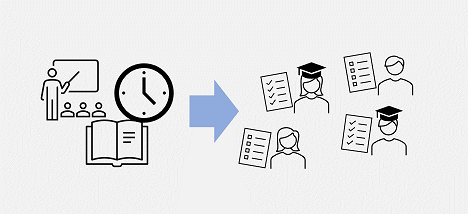
Depois de décadas de desenvolvimento de software livre, a Python fornece funcionalidades abrangentes com bibliotecas estatísticas e numéricas poderosas:

* NumPy e Pandas simplificam a análise e a manipulação de dados
* Matplotlib fornece visualizações de dados atraentes
* Scikit-learn proporciona análises de dados preditivas simples e eficazes
* TensorFlow e PyTorch fornecem recursos de aprendizado profundo e aprendizado de máquina

**Cenário de exemplo**

Normalmente, um projeto de análise de dados é projetado para estabelecer insights sobre um cenário específico ou para testar uma hipótese.

Por exemplo, suponha que um professor universitário colete dados de seus alunos, incluindo o número de palestras assistidas, as horas gastas nos estudos e a nota final do exame do final do semestre. O professor pode analisar os dados para determinar se há uma relação entre a quantidade de estudo que um aluno realiza e a nota final obtida. O professor pode usar os dados para testar uma hipótese de que somente os alunos que estudarem por um número mínimo de horas podem esperar obter uma nota para passar de ano.



**O que faremos?**

Neste módulo de treinamento, exploraremos e analisaremos dados das notas de uma turma universitária fictícia do ponto de vista de um professor. Usaremos notebooks Jupyter e várias ferramentas e bibliotecas do Python para limpar o conjunto de dados, aplicar técnicas estatísticas para testar várias hipóteses sobre os dados e a fim de visualizar os dados para determinar as relações entre variáveis.

**Explorar dados com NumPy e Pandas**

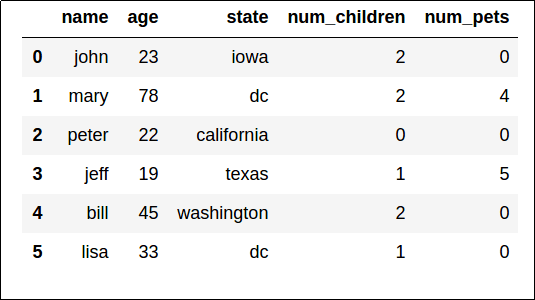
Os cientistas de dados podem usar diversas ferramentas e técnicas para explorar, visualizar e manipular dados. Uma das formas mais comuns de os cientistas de dados trabalharem com os dados é usar a linguagem Python e alguns pacotes específicos para o processamento de dados.

**O que é NumPy?**

NumPy é uma biblioteca do Python que fornece uma funcionalidade comparável a ferramentas matemáticas como MATLAB e R. Embora NumPy simplifique significativamente a experiência do usuário, também proporciona funções matemáticas abrangentes.

**O que é Pandas?**

Pandas é uma biblioteca Python extremamente popular para análise e manipulação de dados. O Pandas é como um aplicativo de planilha para o Python e fornece funcionalidades fáceis de usar para tabelas de dados.



**Explorar dados em um Jupyter Notebook**

Os notebooks Jupyter são uma maneira popular de executar scripts básicos usando o navegador da Web. Normalmente, esses notebooks estão uma única página da Web, dividida em seções de texto e seções de códigos executados no servidor e não no computador local. Executar o código em notebooks Jupyter em um servidor possibilita que você comece rapidamente sem precisar instalar o Python ou outras ferramentas no computador local.

**Teste de hipóteses**

A exploração e a análise de dados normalmente são um processo *iterativo*, no qual o cientista de dados usa uma amostra dos dados e executa os seguintes tipos de tarefa para analisá-la e testar as hipóteses:

* **Limpar dados** para lidar com erros, valores ausentes e outros problemas.
* **Aplicar técnicas estatísticas para entender melhor os dados** e como pode ser esperado que a amostra represente a população do mundo real dos dados, permitindo variação aleatória.
* **Visualizar os dados** para determinar as relações entre as variáveis e, no caso de um projeto de aprendizado de máquina, identificar os *recursos* que sejam potencialmente previsíveis do *rótulo*.
* **Analisar a hipótese** e repetir o processo.

**Visualizar dados**

Os cientistas de dados visualizam os dados para compreendê-los melhor. Eles podem verificar os dados brutos, examinar medidas resumidas, como médias, ou representar graficamente os dados. Os grafos são um meio poderoso de visualizar dados, e os cientistas de dados geralmente usam grafos para discernir padrões moderadamente complexos rapidamente.

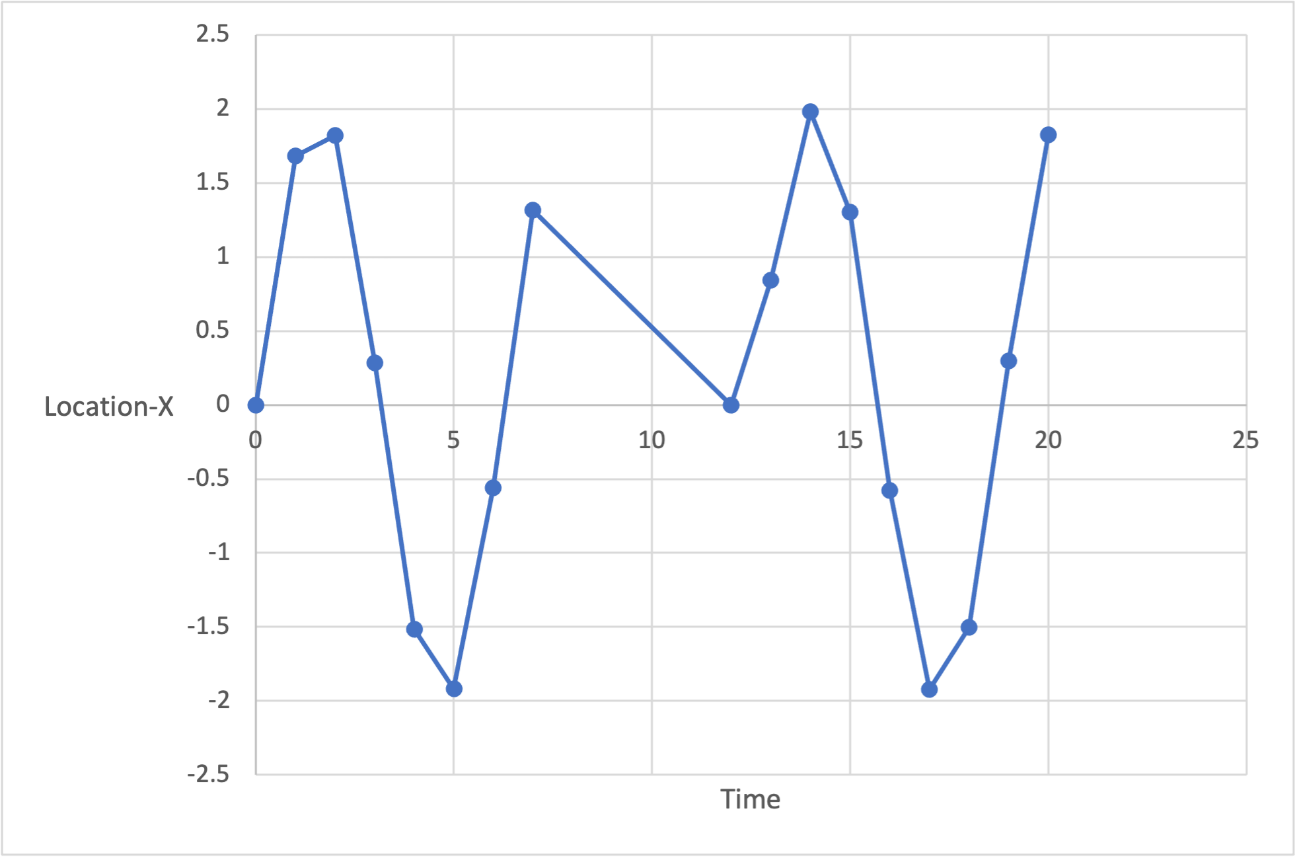
**Representação visual dos dados**

A geração de grafos é feita para fornecer uma avaliação qualitativa rápida de nossos dados, o que pode ser útil para entender os resultados, localizar valores de exceção, verificar como os números são distribuídos e assim por diante.

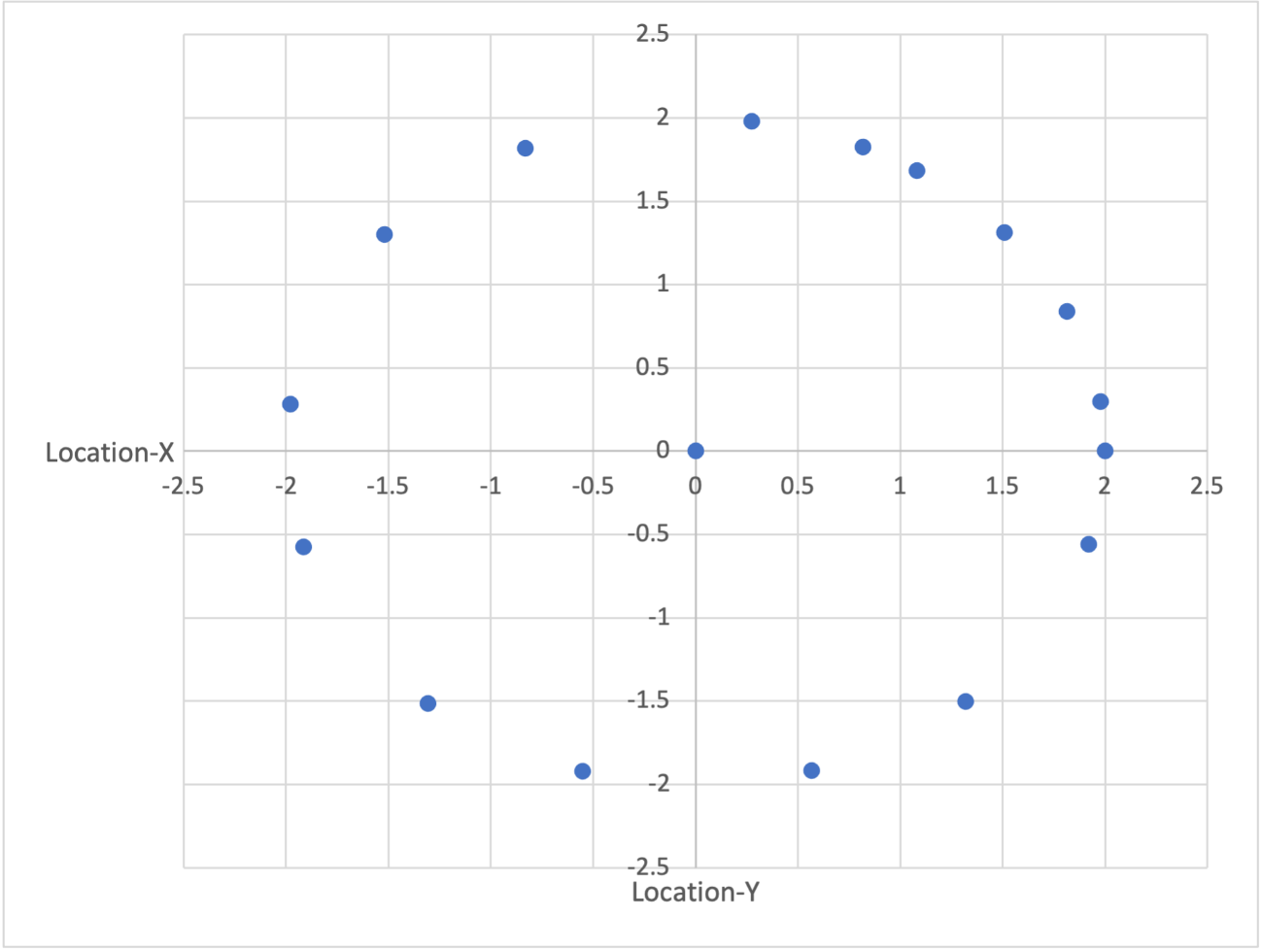
Embora, às vezes, saibamos antecipadamente qual tipo de grafo será mais útil, outras vezes usamos grafos de forma exploratória. Para entender o poder da visualização de dados, considere os seguintes dados: a localização (x, y) de um carro autônomo. É difícil ver padrões reais na forma bruta dos dados. O significado ou a média nos informa que o caminho do carro foi centralizado em torno de x=0,2 e y=0,3, e o intervalo de números parece estar entre cerca de -2 e 2.

| **Hora** | **Local-X** | **Local-Y** |
| --- | --- | --- |
| 0 | 0 | 2 |
| 1 | 1,682942 | 1,080605 |
| 2 | 1,818595 | -0,83229 |
| 3 | 0,28224 | -1,97998 |
| 4 | -1,5136 | -1,30729 |
| 5 | -1,91785 | 0,567324 |
| 6 | -0,55883 | 1,920341 |
| 7 | 1,313973 | 1,507805 |
| 12 | 0,00001 | 0,00001 |
| 13 | 0,840334 | 1,814894 |
| 14 | 1,981215 | 0,273474 |
| 15 | 1,300576 | -1,51938 |
| 16 | -0,57581 | -1,91532 |
| 17 | -1,92279 | -0,55033 |
| 18 | -1,50197 | 1,320633 |
| 19 | 0,299754 | 1,977409 |
| 20 | 1,825891 | 0,816164 |

Se agora plotarmos o Local-X ao longo do tempo, poderemos ver que parece haver alguns valores ausentes entre 7 e 12 vezes.



Se grafarmos X versus Y, teremos um mapa de onde o carro passou. Fica óbvio que o carro estava em um círculo, e em algum momento ele foi até o centro desse círculo.



Os grafos não estão limitados a gráficos de dispersão 2D como os encontrados acima. Eles podem ser usados para explorar outros aspectos dos seus dados, por exemplo, proporções (gráficos de pizza e gráficos de barras empilhadas) e como os dados são distribuídos (histogramas e diagramas de caixa). Geralmente, quando estamos tentando entender os dados brutos ou os resultados, podemos experimentar diferentes tipos de grafos até encontrarmos um que explique os dados de maneira visualmente intuitiva.