Aula 1A - Introdução à Ciência de Dados

Gustavo Oliveira[1] e Andrea Rocha[1]

[1]Departamento de Computação Científica / UFPB

Junho de 2020

1 Introdução à Ciência de Dados

1.1 Ciência de dados no século XXI

A dinâmica do mundo globalizado elevou a importância dos dados e da informação a uma escala jamais vista na história humana em virtude da evolução exponencial dos recursos tecnológicos, dos meios de comunicação e, principalmente, da computação de alto desempenho. Está em vigor a *Era da Informação*, em que os dados são considerados a matéria-prima imprescindível.

Assim como a terra era o recurso fundamental para a agricultura, e o ferro o era para a indústria, os dados tornaram-se um bem de valor inestimável para pessoas, empresas, governos e para a própria ciência. Com a expansão do fenômeno *Big Data*, diversos nichos do conhecimento começaram a eclodir trazendo consigo uma série de nomes elegantes, tais como *business intelligence*, *data analytics*, *data warehouse* e *data engineering*.

Apesar disso, *ciência de dados* desponta-se como o conceito mais razoável para denotar o aspecto científico dos dados. Em um contexto acadêmico, ela encontra-se na interseção de outras áreas do conhecimento e no cerne de uma cadeia maior envolvendo gestão de processos e o pensamento científico.

É difícil estabelecer um modelo holístico unificado que traduza de maneira exata a capilaridade da ciência de dados nas atividades modernas. Diante disso, a Figura 1 tenta ilustrar, para nossos propósitos, como a ciência de dados relaciona-se com outros domínios do conhecimento de maneira multidisciplinar.

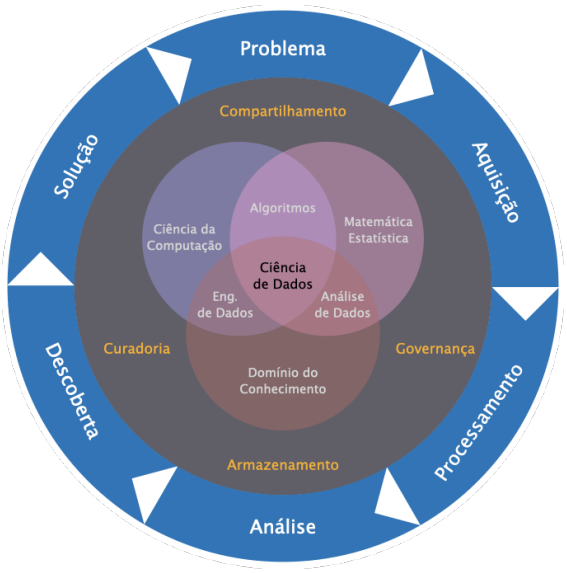

Figura 1: Diagrama com visão holística da Ciência de Dados e áreas correlatas.

O diagrama possui três camadas. A camada mais interna mostra como áreas do conhecimento tradicionais se intersectam para dar forma ao que chamamos de *ciência de dados*. Aqui, enfatizamos três grandes conjuntos:

1. **Matemática/Estatística**, que fornece os modelos matemáticos e estatísticos fundamentais para estudo, análise e inferência de dados, aos quais se agregam as técnicas de aprendizagem de máquina;
2. **Ciência da Computação/Engenharia de Software**, que fornece elementos básicos de hardware e software para projetar soluções de intercâmbio, armazenamento e segurança de dados, por exemplo.
3. **Conhecimento do Domínio/Expertise**, que é o próprio ramo de aplicação do conhecimento que está sendo buscado através dos dados em questão, ao qual se aderem o *data reporting*, a inteligência de negócios, o *marketing* e a comunicação de dados em geral para suporte à tomada de decisões.

A camada intermediária relaciona-se à gestão de processos da cadeia de dados, que envolvem governança, curadoria, armazenamento e reuso de dados, por exemplo, isto é, todos os aspectos relacionados à preservação, manutenção, destruição e compartilhamento de dados.

No invólucro mais externo, temos a camada relativa ao método científico de busca de soluções para um dado problema. Com alguma adaptação, os processos envolvidos nesta camada representam de maneira satisfatória tanto a ideia de *soluções dirigidas por dados* (*data-driven solutions*) amplamente utilizada em contextos empresariais e industriais, em que ferramentas inovadoras são construídas para entregar produtos e soluções especialmente voltadas a um segmento ou público particular com base em um cuidadoso mapeamento de clientes, quanto o compartilhamento e reprodutibilidade da pesquisa científica. Em linhas gerais, este ciclo contém os seguintes processos:

1. **Definição do problema**, etapa em que uma "grande pergunta" é feita, a qual, a princípio, pode ser respondida ao se vasculhar um conjunto de dados específico.

2. **Aquisição de dados**, etapa em que se coleta toda a informação relacionada ao problema lançado na etapa anterior.
3. **Processamento de dados**, etapa em que os dados adquiridos são processados para análise. Nesta etapa realiza-se um verdadeiro tratamento dos dados (limpeza, formatação e organização).
4. **Análise de dados**, etapa em que os dados são analisados e perscrutados por meio de técnicas de mineração, agrupamento e clusterização. Neste momento é que testes de hipótese e mecanismos de inferência são utilizados.
5. **Descoberta de dados**, etapa em que descobertas são realizadas, tais como correlações entre variáveis, comportamentos distintivos e tendências claramente identificáveis, permitindo que conhecimento seja gerado a partir da informação.
6. **Solução**, etapa final do ciclo na qual as descobertas podem ser convertidas em produtos e ativos de valor agregado para o domínio do problema proposto.

1.1.1 O caso da COVID-19

A pandemia causada pela COVID-19 que assolou o mundo recentemente pode ser tomada como um estudo de caso singular de aplicação do processo de análise de dados citado na seção anterior. Sob o ponto de vista científico, poderíamos levantar várias questões acerca do vírus no que diz respeito à velocidade de contágio, ao impacto em atividades econômicas, às alterações no comportamento social, entre outras.

Modelos epidemiológicos apontam que a interação entre pessoas é um dos principais mecanismos de transmissão viral. A partir dessa premissa e levando em consideração o nosso país, uma pergunta que poderíamos fazer a fim de nortear uma pesquisa em ciência de dados seria: *a taxa de contágio do vírus em pessoas vivendo próximas de um centro comercial localizado em uma zona rural é menor do do que em pessoas vivendo próximas de um centro comercial localizado em uma zona urbana?*. É evidente que, para responder uma pergunta como esta com precisão científica, necessitaríamos de definições e muitos dados. Como delimitaríamos a zona urbana? O centro comercial deveria ser caracterizado como um conjunto de lojas de pequeno porte? Feiras? Um local de comércio onde, diariamente, circulam 100 pessoas por hora? Além disso, neste caso, como faríamos para coletar as informações de que precisamos? No banco de dados do IBGE? No DATASUS?

A aquisição de dados pode ser uma tarefa mais difícil do que se imagina. No caso em questão, certamente buscaríamos informações em bancos de dados do setor público, de secretarias municipais, de órgãos estaduais, até instituições especializadas em âmbito federal. Entretanto, no caso do Brasil, nem todas as regiões – quiçá o país inteiro – usufruem de bancos de dados amplos e precisos onde variáveis primárias necessárias para a análise de dados sejam facilmente obtidas.

Supondo que tenhamos em mãos as informações de saúde acerca dos habitantes das zonas rural e urbana necessárias para nossa pesquisa sobre a COVID-19, o outro passo a tomar é o processamento dos dados. De que maneira o banco de dados se apresenta? Como uma infinidade de planilhas de Excel sem nenhuma formatação específica? Arquivos .csv estruturados e categorizados por faixa etária, município, densidade populacional? Toda a informação é hierárquica em arquivos HDF5?

Para cada situação, devemos dispor de ferramentas específicas e adequadas para manipular, organizar, limpar e estruturar os dados. Todo este tratamento dos dados ocorre, em geral, por duas vias: soluções pré-existentes (programas, recursos, interfaces, frameworks, projetos *open source* etc.

já disponíveis no mercado ou na academia) ou soluções customizadas, criadas pelo cientista de dados para o atendimento de demandas específicas não cobertas pelas soluções pré-existentes.

Uma vez processados, os dados atingem uma condição minimamente razoável para serem escrutinados, isto é, analisados minuciosamente. Nesta fase, o intelecto de quem analisa os dados está a todo vapor, visto que um misto de conhecimento técnico, experiência, e criatividade são os ingredientes para realizar descobertas. Os dados são levados de um lado para outro, calculam-se expressões matemática aqui e acolá, testes estatísticos são feitos uma, duas, três, n vezes, até que conclusões surpreendentes podem aparecer.

A propagação de um vírus é um fenômeno não linear suscetível a dinâmicas quase imprevisíveis. Portanto, ao procurarmos a resposta para uma pergunta difícil como a que pusemos acima, pode ser que descubramos padrões e tendências que sequer cogitávamos capazes de responder até mesmo perguntas para outros problemas. Poderíamos chegar à conclusão, por exemplo, que a taxa de contágio na zona urbana é afetada pelas características arquitetônicas do centro comercial: arejamento deficiente, corredores de movimentação estreitos, pontos de venda altamente concentrados, etc.

Ao final do ciclo, espera-se que respostas sejam obtidas para que soluções sejam propostas e decisões tomadas com responsabilidade. Quando o assunto é a saúde de pessoas, questões éticas e morais tornam-se extremamente sensíveis. O papel de cientistas e analistas de dados em situações particulares como a da COVID-19 é munir gestores e líderes com recomendações resultantes das evidências mostradas pelos dados. Todavia, é importante dizer que modelos matemáticos são estimativas da realidade e também possuem graus de falibilidade. Portanto, equilibrar as descobertas com o peso das decisões é essencial para o alcance de soluções adequadas.

Diversos projetos focados em ciência e análise de dados focados no estudo da COVID-19 estão atualmente em curso no mundo. Um dos pioneiros foi o *Coronavirus Resource Center* da *John Hopkins University* [CRC-JHU]. Iniciativas no Brasil são as seguintes: *Observatório Covid-19 BR* [COVID19BR], *Observatório Covid-19 Fiocruz* [FIOCRUZ], CoronaVIS-UFRGS [CoronaVIS-UFRGS], CovidBR-UFCG [CovidBR-UFCG], entre outras. Na UFPB, destacamos a página do LEAPIG [LEAPIG-UFPB]. Certamente, a COVID-19 deverá se consagrar como um dos maiores estudos de caso da história mundial para a ciência e análise de dados, haja vista o poder computacional de nossos dias.

1.1.2 Cientista de dados x analista de dados x engenheiro de dados

As carreiras profissionais neste novo mundo dos dados converteram-se em muitas especialidades. Há três perfis, em particular, sobre os quais gostaríamos de comentar: *o cientista de dados*, o *analista de dados* e o *engenheiro de dados*. Porém, antes de entrar nesta "sopa de letrinhas", vale a pena entender um pouco sobre como a ciência de dados, como um todo, é compreendida pelas pessoas mundo afora.

Nos Estados Unidos, um esforço conjunto entre representantes da universidade, do poder público, da indústria e de outros segmentos culminou na publicação especial No. 1500-1 (2015) do *National Institute of Standards and Technology* (NIST), que definiu diversos conceitos relacionados à ciência de dados [NIST 1500-1 (2015)]. Segundo este documento,

> *"**Cientista de dados** é um profissional que tem conhecimentos suficientes sobre necessidades de negócio, domínio do conhecimento, além de possuir habilidades analíticas, de software e de engenharia de sistemas para gerir, de ponta a ponta, os processos envolvidos no ciclo de vida dos dados."*

Como se vê, a identidade do cientista de dados é definida por uma interseção de competências. Todas essas competências estão distribuídas, de certa forma, nas três grandes áreas do conhecimento que citamos acima. Por outro lado, o que exatamente é a *ciência de dados*?

De acordo com o mesmo documento,

> *"**Ciência de dados** é a extração do conhecimento útil diretamente a partir de dados através de um processo de descoberta ou de formulação e teste de hipóteses."*

A partir disso, percebemos que *informação* não é sinônimo de *conhecimento*. Para termos uma clareza melhor dessa distinção, basta refletirmos sobre nosso uso diário do celular. O número de mensagens de texto, de fotos, áudios e vídeos que trocamos com outras pessoas por meio de aplicativos de mensagem instantânea, redes sociais ou e-mail é gigantesco. Quantos de nós não passamos pela necessidade de apagar conteúdo salvo em nosso celular para liberar espaço! Às vezes, não temos ideia de quanta informação trocamos por minuto com três ou quatro colegas. A questão central é: de toda essa informação, que fração seria considerada útil? Isto é, o que poderíamos julgar como conhecimento aproveitável? A resposta talvez seja um incrível "nada"...

Portanto, ter bastante informação à disposição não significa, necessariamente, possuir conhecimento. Da mesma forma que estudar para aprender é um exercício difícil para o cérebro, garimpar conhecimento em meio a um mar de informação é uma tarefa que exige paciência, análise, raciocínio dedutivo e criatividade. Por falar em análise de dados, vamos entender um pouco sobre o termo *analytics*, frequentemente utilizado no mercado de trabalho.

Analytics pode ser traduzido literalmente como "análise" e, segundo o documento NIST 1500-1, é definido como o "processo de sintetizar conhecimento a partir da informação". Diante disso, podemos dizer que

> *"**Analista de dados** é o profissional capaz de sintetizar conhecimento a partir da informação e convertê-lo em ativo exploráveis."*

Uma terceira vertente que surgiu com a evolução do *Big Data* foi a *engenharia de dados*, que tem por objetivo projetar ferramentas, dispositivos e sistemas com robustez suficiente para lidar com a grande massa de dados em circulação. Podemos dizer que

> *"**Engenheiro(a) de dados** é o(a) profissional que explora recursos independentes para construir sistemas escaláveis capazes de armazenar, manipular e analisar dados com eficiência e e desenvolver novas arquiteturas sempre que a natureza do banco de dados exigi-las."*

Embora essas três especializações possuam características distintivas, elas são tratadas como partes de um corpo maior, que é a Ciência de Dados. O projeto EDISON, coordenado pela Universidade de Amsterdã, Holanda, por exemplo, foi responsável por mapear definições e taxonomias para construir grupos profissionais em ciência de dados para ocuparem posições em centros de pesquisa e indústrias na Europa. De acordo com o *EDISON Data Science Framework* [EDSF], os grupos profissionais se dividem entre gerentes (CEOs, líderes de pesquisa), profissionais gerais (analista de negócios, engenheiros de dados etc.), profissionais de banco de dados (designer de computação em nuvem, designer de banco de dados etc.), profissionais de curadoria (bibliotecários, arquivistas etc.), profissionais técnicos (operadores de equipamentos, mantenedores de *warehouses* etc.) e profissionais de apoio (suporte a usuários, alimentadores de sistemas, atendentes etc.).

Quem faz o quê? Resumimos a seguir as principais tarefas atribuídas a cientistas, analistas e engenheiros(as) de dados com base em artigos de canais especializados [DataQuest], [NCube],

[Medium], [Data Science Academy], [Data Flair]. Uma característica importante entre os perfis diz respeito à organização dos dados. Enquanto cientistas e analistas de dados lidam com dados *estruturados* – dados organizados e bem definidos que permitem fácil pesquisa –, engenheiros(as) de dados trabalham com dados *não estruturados*.

Cientista de dados

- Realiza o pré-processamento, a transformação e a limpeza dos dados;
- Usa ferramentas de aprendizagem de máquina para descobrir padrões nos dados;
- Aperfeiçoa e otimiza algoritmos de aprendizagem de máquina;
- Formula questões de pesquisa com base em requisitos do domínio do conhecimento;

Analista de dados

- Analisa dados por meio de estatística descritiva;
- Usa linguagens de consulta a banco de dados para recuperar e manipular a informação;
- Confecciona relatórios usando visualização de dados;
- Participa do processo de entendimento de negócios;

Engenheiro(a) de dados

- Desenvolve, constroi e mantém arquiteturas de dados;
- Realiza testes de larga escala em plataformas de dados;
- Manipula dados brutos e não estruturados;
- Desenvolve *pipelines* para modelagem, mineração e produção de dados
- Cuida do suporte a cientistas e analistas de dados;

Que ferramentas são usadas? As ferramentas usadas por cada um desses profissionais são variadas e evoluem constantemente. Na lista a seguir, citamos algumas.

Cientista de dados

- R, Python, Hadoop, Ferramentas SQL (Oracle, PostgreSQL, MySQL etc.)
- Álgebra, Estatística, Aprendizagem de Máquina
- Ferramentas de visualização de dados

Analista de dados

- R, Python,
- Excel, Pandas
- Ferramentas de visualização de dados (Tableau, Infogram, PowerBi etc.)
- Ferramentas para relatoria e comunicação

Engenheiro(a) de dados

- Ferramentas SQL e noSQL (Oracle NoSQL, MongoDB, Cassandra etc.)
- Soluções ETL - Extract/Transform/Load (AWS Glue, xPlenty, Stitch etc.)
- Python, Scala, Java etc.
- Spark, Hadoop etc.

1.1.3 Matemática por trás dos dados

No mundo real, lidamos com uma grande diversidade de dados, mas nem sempre percebemos como a Matemática atua por trás de cada pedacinho da informação. Ao longo da sua graduação em ciência de dados, você aprenderá conceitos abstratos novos e trabalhará com mais profundidade outros que já conhece, tais como vetor e matriz.

Você provavelmente já deve ter ouvido falar que o computador digital funciona com uma linguagem *binária* cujas mensagens são todas codificadas como sequencias dos dígitos 0 e 1. Daí que vem o nome *bit*, um acrônimo para *binary digit*, ou dígito binário.

Em termos de *bits*, a frase "Ciência de dados é legal!", por exemplo, é escrita como

`100001111010011110101011011101100011110100111000011000001100100110010100001100100110000111001001101111111001110000011101001100000110110011001101111100001110100100001.`

Interessante, não? Vejamos outros exemplos.

Nas aulas de Física, você aprendeu que um vetor possui uma origem e uma extremidade. Tanto a origem como a extremidade são "pontos" do espaço. Agora, imagine um plano cartesiano. Se a sua origem é o ponto $O = (0, 0)$ e a sua extremidade é o ponto $B = (2, 0)$, você pode traçar um vetor de O a B andando duas unidades para a direita. Claramente, este vetor estará sobre o eixo das abscissas. Imagine que você pudesse então usar cada unidade como se fosse uma "caixa" onde pudesse "guardar" uma informação sobre você. Ou seja, em $O = (0, 0)$ você coloca seu nome, em $A = (0, 1)$ a sua idade e em $B = (0, 2)$ seu CPF. Você teria um "vetor" com 3 valores. Além disso, suponha que você pudesse fazer o mesmo para mais 9 pessoas de sua família repetindo este processo em outros 9 vetores paralelos ao primeiro. Você teria agora $3 + 9 \times 3 = 3 + 27 = 30$ caixas para guardar informações. OK, e daí?

O que acabamos de ilustrar é a ideia fundamental para estruturar tabelas, planilhas do Excel, ou *dataframes* (que você aprenderá neste curso). Tudo isso são matrizes! Ou seja, informação organizada em linhas e colunas! Cada linha é como um vetor que contém 3 posições (são as colunas). Cada coluna são os registros que você coloca. Então, digamos que você tenha pensado na sua mãe como o próximo membro da família. O nome dela seria colocado na "caixa" que estaria no ponto $(1, 0)$, a idade dela na "caixa" que estaria no ponto $(1, 1)$ e o CPF dela na "caixa" que estaria no ponto $(1, 2)$. Fazendo o com todos os demais membros, você vai concluiria que o CPF do 10o. membro da família deveria estar na "caixa" associada ao ponto $(9, 9)$.

Para um computador, vetores são chamados de *arrays*. Uma lista de coisas também pode ser comparada a um *array*. Note acima que a segunda coordenada do primeiro vetor (aquele que tem as caixas de informação a seu respeito) é sempre zero. Ela se mantém fixa. Isto significa que o dado assemelha-se a algo **unidimensional**. Isto é, basta que eu apenas faça uma contagem de elementos em uma direção. Isto é muito similar ao conjunto dos números inteiros positivos $\mathbb{Z}_+$.

Quando, porém, inserimos os vetores adicionais (as informações dos membros da sua família), a segunda coordenada também se altera. Isto significa que o dado assemelha-se a algo **bidimensional**. Ou seja, a contagem dos elementos ocorre em duas direções. Levando em conta um plano cartesiano com o eixo das ordenadas orientado para baixo e não para cima, como de costume, os números cresceriam não apenas para a direita, mas também para baixo. Logo, teríamos uma segunda contagem baseada em mais um conjunto de números inteiros positivos $\mathbb{Z}_+$ independente do primeiro. De que estamos falando aqui? Estamos falando do conceito de *par ordenado*. Isto é,

qualquer ponto (x, y) com $x \in \mathbb{Z}_+$ e $y \in \mathbb{Z}_+$ é um "local" onde existiria uma caixinha onde podemos guardar informações de maneira independente. Uma matriz é exatamente isto.

As imagens vistas na televisão, as *selfies* e fotografias que você faz com seu celular e as figuras neste livro podem todas ser descritas como matrizes. Cada elemento da matriz é identificado com uma posição (x, y) ao qual damos o nome de *pixel*. Uma imagem é, por sua vez, uma pixelização. Porém, as imagens não são apenas "endereços" de pixels. Elas possuem cor, tons de cinza, ou são monocromáticas (preto e branco). As cores são representadas por "canais". E, acredite, cada canal é também uma matriz de dados! No final das contas, uma imagem colorida é uma "matriz formada por outras matrizes"!

Uma matriz formada a partir de outra matriz é um exemplo de dado **tridimensional**. Um exemplo disso são dados sequenciais, tais como um filme ou uma animação. O número de *frames per second* (FPS), ou "quadros por segundo", é tão alta hoje em dia que nossa visão não é capaz de captar que, quando vamos ao cinema ou assistimos um filme pela TV ou no Youtube, o que vemos é exatamente a mudança rápida e sucessiva de vários "quadros" de imagens por segundo.

Como você verá ao longo deste curso, muitos conceitos de Matemática que você aprendeu ao longo do Ensino Médio começarão a fazer mais sentido com as aplicações.

1.2 Ferramentas computacionais do curso

Neste curso, usaremos Python 3.x (onde x é um número de versão) como linguagem de programação. Por se tratar de uma linguagem interpretada, interagir com ela é mais fácil do que uma linguagem compilada. Um conjunto mínimo de recursos para Python funcionar é composto do *core* da linguagem, um terminal de comandos e um editor de texto. Enquanto programadores experientes usam menos recursos visuais, para efeito didático, usaremos interfaces mais amigáveis e interativas comprovadas como bons ambientes de aprendizagem.

1.2.1 *iPython* e *Jupyter Notebook*

O [iPython] foi um projeto iniciado em 2001 para o desenvolvimento de um interpretador Python para melhorar a interatividade com a linguagem. Ele foi integrado como um *kernel* (núcleo) no projeto [Jupyter], desenvolvido em 2014, permitindo textos, códigos e elementos gráficos sejam integrados em cadernos interativos. *Jupyter notebooks* são interfaces onde podemos executar códigos em diferentes linguagens desde que alteremos os *kernels*. A palavra *Jupyter* é uma aglutinação das iniciais de *Julia*, *Python* e *R*, que são as linguagens de programação mais usuais para ciência de dados.

1.2.2 *Anaconda*

Em 2012, o projeto [Anaconda] foi iniciado como objetivo de fornecer uma ferramenta completa para o trabalho com Python. Em 2020, já como uma empresa de ponta, ela tornou-se uma das pioneiras no fornecimento de plataformas individuais e empresariais para ciência de dados. Segundo a empresa, a [Individual Edition], que é a versão aberta para uso é a mais popular no mundo com mais de 20 milhões de usuários. Recomendamos que você siga as orientações de instalação desta versão. Uma vez instalada, basta lançar as ferramentas a partir do dashboard *Anaconda Navigator*.

1.2.3 *Jupyter Lab*

Uma ferramenta que melhorou a interatividade do Jupyter é o *Jupyter Lab*, que realiza um alto nível de integração. Este [artigo] discute as características do Jupyter Lab, entre as quais vale citar o recurso de arrastar/soltar para reordenar células de cadernos e copiá-las entre cadernos.

1.2.4 *Binder*

O projeto [Binder] funciona como um servidor online baseada na tecnologia *Jupyter Hub* para servir cadernos interativos online. Através do Binder, é possível executar códigos "na nuvem" sem a necessidade de instalações, porém as sessões são perdidas após o uso.

1.2.5 *Google Colab*

O [Google Colab], uma redução de *Colaboratory*, é uma ferramenta que possui características mistas entre o *Jupyter notebook* e o *Binder*, porém permite que o usuário use a infra-estrutura de computação de alto desempenho (GPUs e TPUS) da Google. A vantagem é que usuários de contas Google podem sincronizar arquivos diretamente com o Google Drive.

1.2.6 Módulos principais

Neste curso, o ecossistema de ferramentas torna-se pleno com a adição de alguns módulos que são considerados essenciais para a prática da ciência e análise de dados contemporânea:

- *numpy* (*NUMeric PYthon*): o *numpy* serve para o trabalho de computação numérica, operando fundamentalmente com vetores, matrizes e ágebra linear.
- *pandas* (*Python for Data Analysis*): é a biblioteca para análise de dados de Python, que opera *dataframes* com eficiência.
- *sympy* (*SYMbolic PYthon*): é um módulo para trabalhar com matemática simbólica e cumpre o papel de um verdadeiro sistema algébrico computacional.
- *matplotlib*: voltado para plotagem e visualização de dados, foi um dos primeiros módulos Python para este fim.
- *scipy* (*SCIentific PYthon*): o *scipy* pode ser visto, na verdade, como um módulo mais amplo que integra os módulos anteriores. Em particular, ele é utilizado para cálculos de integração numérica, interpolação, otimização e estatística.
- *seaborn*: é um módulo para visualização de dados baseado no *matplotlib*, porém com capacidades visuais melhores.

A visualização de dados é um tema de suma importância para resultados da análise exploratória de dados em estatística. Um site recomendado para pesquisar as melhores ferramentas para análise de dados é o [PyViz].