Uma introdução à análise de dados usando pandas, matplotlib e seaborn

Vitor Moreira Casagrande[1], Thiago Pereira da Silva[1]

[1]Campus Universitário do Araguaia – Universidade Federal de Mato Grosso (UFMT)
Barra do Garças – MT – Brazil

vitormoreiracasagrande@hotmail.com.br, thiago.silva@ufmt.br

Abstract. *With accelerated digitalization, Data Science has evolved by combining statistics and programming, enabling complex analyses of large data volumes. Tools like Python and R, along with specialized libraries, have facilitated the use of data analysis techniques and the construction of predictive models applicable to diverse fields, such as education and healthcare. In education, data analysis is used to personalize learning and identify students at risk of dropout, while in healthcare, it was essential for monitoring and containing COVID-19. This practical course provides an introduction to these tools and methods, empowering students to conduct initial analyses and communicate valuable insights.*

Resumo. *Com a digitalização acelerada, a Ciência de Dados evoluiu ao combinar estatística e programação, permitindo análises complexas de grandes volumes de dados. Ferramentas como Python e R, junto a bibliotecas específicas, facilitam o uso de técnicas de análise de dados e a construção de modelos preditivos aplicáveis em diversas áreas, como por exemplo, educação e saúde. Na educação, a análise de dados é usada para personalizar o ensino e identificar alunos em risco de evasão, enquanto na saúde foi fundamental para monitorar e conter a Covid-19. Esta oficina oferece uma introdução a essas ferramentas e métodos, capacitando os alunos a realizar análises iniciais e comunicar insights relevantes.*

1. Introdução

Com as recorrentes transformações digitais, diversos aspectos da vida humana passaram a ser registrados em formato digital, resultando em um crescimento exponencial do volume de dados gerados diariamente. Dados que antes eram coletados e armazenados de forma analógica agora estão digitalizados, possibilitando um potencial maior para análise e compreensão. A integração de técnicas de programação e software ao campo da Estatística deu origem ao que chamamos de Ciência de Dados (do inglês *Data Science*), que permite extrair estatísticas, gerar insights e até realizar previsões, independentemente da quantidade de dados. Após anos de evolução e ampla adoção no meio acadêmico e corporativo, a Ciência de Dados se expandiu, criando carreiras especializadas como Análise de Dados, Engenharia de Dados e Engenharia de Machine Learning. Esses campos frequentemente se sobrepõem, formando uma cadeia que abrange desde a coleta de dados até a aplicação de modelos preditivos e a implementação de soluções baseadas em *insights* obtidos [Simplilearn 2024].

O aumento dos dados digitais e o desenvolvimento de tecnologias de Big Data permitiram análises mais detalhadas e sofisticadas de grandes volumes de informações, ampliando as possibilidades em relação às ferramentas estatísticas tradicionais antes usadas para cálculos e testes de hipóteses. Conforme [Zhang, Wolfram e Ma, 2023], essas novas tecnologias possibilitam o processamento de dados em escalas e velocidades antes inimagináveis, oferecendo uma base mais robusta para modelos preditivos e insights complexos que sustentam decisões estratégicas.

A evolução dos dados foi acompanhada pela criação de linguagens de programação como Python e R, que, juntamente com bibliotecas específicas como Pandas[1], Matplotlib[2], Seaborn[3] e Scikit-Learn,[4] impulsionaram a análise de dados e facilitaram a aplicação de técnicas de Machine Learning. Essas ferramentas não apenas tornaram possível a construção de modelos preditivos, mas também permitiram a realização de análises profundas e abrangentes, proporcionando *insights* valiosos a partir dos dados. Essa combinação de linguagens e bibliotecas impulsionou o desenvolvimento de soluções inovadoras em diversas áreas, otimizando processos e melhorando a tomada de decisões.

A análise de dados tem um grande impacto em diversas áreas do nosso cotidiano, como saúde e educação. Na educação, instituições de ensino utilizam dados de desempenho acadêmico dos alunos para identificar perfis de aprendizado e personalizar o conteúdo trabalhado em sala de aula. Com essa abordagem, é possível identificar estudantes que precisam de suporte adicional ou ajustes em atividades e avaliações, promovendo um ensino mais inclusivo e eficaz. Outro desafio enfrentado pela educação no Brasil é a alta taxa de evasão escolar, que afeta tanto instituições públicas quanto privadas. Como apontado por [Romero e Ventura, 2010], uma das formas de combater a evasão é identificar precocemente os perfis de alunos com maior risco de abandono. Com esses dados, gestores educacionais podem desenvolver estratégias e políticas de retenção mais assertivas, prevenindo o abandono escolar e promovendo a continuidade dos estudos.

Na área da saúde, a análise de dados também desempenha um papel crucial na identificação e monitoramento de doenças. Durante a pandemia de Covid-19, por exemplo, os dados de casos positivos permitiram mapear as regiões mais afetadas e orientar ações de contenção específicas. Esse monitoramento em tempo real ajudou os governos a tomar decisões que salvaram vidas, controlaram a propagação do vírus e reduziram o impacto sobre o sistema de saúde, evidenciando o poder dos dados na gestão de crises sanitárias e na proteção da população.

Com o crescimento exponencial da demanda por profissionais capazes de interpretar e extrair valor dos dados, esta oficina oferece uma introdução prática e essencial à análise de dados. Ao longo das aulas, utilizaremos dois datasets amplamente empregados no aprendizado inicial — o MovieLens [Grouplens, 2024] e o TMDB [Keagle, 2024] — que oferecem contextos reais para a aplicação de técnicas de análise. Exploraremos ferramentas e métodos fundamentais, como limpeza e preparação de dados, visualização e técnicas estatísticas básicas, com foco em desenvolver habilidades iniciais de análise e interpretação. Além disso, os participantes aprenderão a utilizar bibliotecas de Python como Pandas, Matplotlib e Seaborn para manipulação e visualização, permitindo uma compreensão superficial dos dados e das tendências que eles revelam. Ao final do curso, os participantes estarão aptos a conduzir análises exploratórias iniciais e a comunicar *insights* de forma clara e objetiva, competências valiosas em qualquer área de atuação.

Este documento está organizado da seguinte forma: A Seção 2 define o termo "análise de dados" e o contextualiza em áreas correlatas. A Seção 3 apresenta de maneira sucinta as principais ferramentas para análise de dados, incluindo uma breve descrição das linguagens de programação e dos ambientes interativos. A Seção 4 detalha o processo estruturado para análise de dados, que consiste em quatro etapas. Por fim, a Seção 5 apresenta as considerações finais.

1 Disponível em: https://pandas.pydata.org/. Acesso em: 30 out. 2024.
2 Disponível em: https://matplotlib.org/. Acesso em: 30 out. 2024.
3 Disponível em: https://seaborn.pydata.org/. Acesso em: 30 out. 2024.
4 Disponível em: https://scikit-learn.org/stable/. Acesso em: 30 out. 2024.

2. Conceituando a análise de dados

A análise de dados é uma disciplina essencial que visa transformar dados brutos em *insights* valiosos e informações práticas, utilizando processos de limpeza, transformação, manipulação e inspeção. Essas etapas são cruciais para apoiar a tomada de decisões informadas [Provost e Fawcett, 2013]. A análise de dados abrange uma ampla gama de técnicas e ferramentas, que variam desde métodos estatísticos básicos, ferramentas de visualização até algoritmos avançados. Além disso, a análise de dados está interligada à ciência de dados, ao Big Data e ao Machine Learning, formando um ecossistema integrado de práticas e tecnologias que possibilitam a extração de valor de grandes volumes de informações, conforme ilustrado na Figura 1. Entender as nuances dessas áreas, assim como suas diferenças, é fundamental para o desenvolvimento de soluções inovadoras para os desafios cotidianos [Zhang, Wolfram e Ma, 2023].

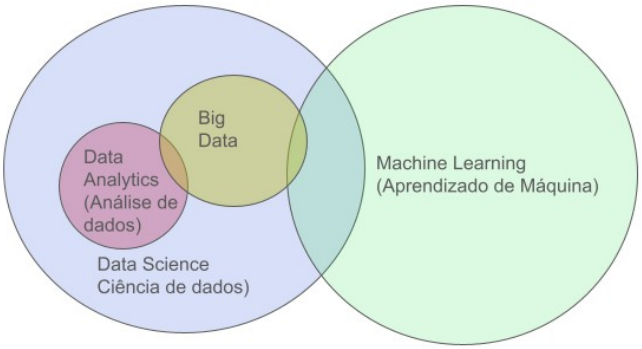

Figura 1 - Conceituando Análise de Dados.
Fonte: Elaborado pelos autores (2024).

A ciência de dados combina programação, estatística e conhecimento especializado para analisar e interpretar dados. Ela envolve a coleta e armazenamento de dados, e o uso de algoritmos complexos, além da criação de modelos preditivos [Provos e Fawcett, 2013]. Por outro lado, Big Data refere-se a conjuntos de dados grandes e complexos que não podem ser processados com ferramentas tradicionais. O conceito Big Data abrange tanto o volume de dados quanto a variedade (tipos de dados) e a velocidade (rapidez com que os dados são gerados e processados), formando o que é conhecido como as três Vs do Big Data [Laney, 2001]. Por fim, Machine Learning é a área que foca no desenvolvimento de algoritmos que permitem que os computadores aprendam com os dados. Os modelos de Machine Learning são empregados para prever resultados, identificar padrões e automatizar decisões com base em dados históricos [Hastie, Tibshirani e Friedman, 2009].

Desta forma, a análise de dados alimenta os modelos de Machine Learning e fornece as bases para a ciência de dados, que, em combinação com técnicas de Machine Learning, é aplicada a conjuntos de dados cada vez maiores e mais complexos, como os gerados no contexto de Big Data [Cavanillas, Curry e Wahlster, 2016].

3. Principais ferramentas para análise de dados

Existem diversas ferramentas utilizadas na análise de dados, que variam em complexidade e aplicabilidade, dependendo das necessidades específicas da análise. Um exemplo são as planilhas eletrônicas, que embora simples, se mostram úteis para a etapa de análise exploratória. As tabelas eletrônicas, em sua maioria, possuem tabelas dinâmicas, funcionalidades para normalização dos dados e gráficos que facilitam a

visualização dos dados. Para análises mais complexas, recorre-se a ferramentas avançadas de Business Intelligence (BI), Machine Learning e Inteligência Artificial, bem como ferramentas de ETL (Extração, Transformação e Carga de Dados, do inglês *Extract, Transform and Load*). No entanto, neste trabalho, focaremos em análises introdutórias, abordando apenas duas categorias de ferramentas, que serão apresentadas a seguir.

3.1. Linguagens de programação

Dentre o conjunto de linguagens de programação para análise de dados, destacam-se as linguagens R e Python. Tais linguagens, com extensões adequadas, oferecem suporte à manipulação dos dados, criação de modelos preditivos e visualizações avançadas. R é uma linguagem e ambiente de software desenvolvido especialmente para estatísticas e é amplamente utilizada em análises complexas e modelagem de dados. Seu uso é mais frequente em análises científicas e contextos acadêmicos.

Python, por outro lado, é uma linguagem de programação de alto nível versátil, de fácil aprendizado e com uma grande variedade de aplicações. A linguagem Python possui várias bibliotecas para análise de dados, como Pandas, NumPy, Matplotlib e Seaborn. A Pandas é a principal biblioteca para manipulação de dados, oferecendo estruturas de dados complexas chamadas de *DataFrames* para organizar dados em formato tabular, além de métodos para seleção, filtragem, agregação, limpeza e transformação de conjunto de dados. A biblioteca NumPy possui um conjunto vasto de operações numéricas e manipulação de *arrays* multidimensionais. Tal biblioteca é base para a biblioteca Pandas.

As bibliotecas Matplotlib e Seaborn possuem funcionalidades para visualização de dados, permitindo a criação desde gráficos simples como gráficos de linha, barras, dispersão, histogramas e mapas de calor, até gráficos complexos como regressão linear, matriz de correlações e gráficos de caixa (*Boxplots*). Esta oficina será focada na linguagem Python, utilizando as bibliotecas supracitadas para manipular dados e explorar visualizações eficazes.

3.2. Ambientes interativos de análise de dados

Atualmente existem dois principais ambientes interativos para análise de dados que oferecem funcionalidades para os usuários manipular e visualizar dados, integrando código, gráficos e texto em um único ambiente. O Jupyter Notebook [Databricks, 2024] é um aplicativo web de código aberto, muito usado para a área de ciência de dados, eles servem para análise exploratória de dados, visualização de dados, Machine Learning, modelagem estatística entre outros. Dentro do Jupyter Notebook, é possível ter vários blocos de código ou texto, facilitando a execução passo a passo de uma análise de dados com *outputs* (saídas) de forma dinâmica. O Jupyter Notebook funciona localmente, o que pode ser um problema caso se trabalhe com um grande volume de dados e as especificações técnicas do computador não suportarem trabalhar com um volume de dados tão grande. Desta forma, por funcionar de forma local, o Jupyter Notebook não é ideal para trabalhos colaborativos.

O Google Colab [Google Colab 2024] é uma plataforma em nuvem, que emula um ambiente de programação que permite executar códigos através de blocos de execução tal qual o Jupyter notebook. Esta ferramenta tem a vantagem de usar os recursos de hardware fornecidos pelo Google, como GPUs (*Graphics Processing Unit*) e TPUs (*Tensor Processing Unit*) para computação acelerada, além de integração com o Google Drive. Em contrapartida, o Google Colab só funciona em modo on-line.

4. Processo estruturado para análise de dados

Tipicamente, a análise de dados constitui um processo estruturado com quatro etapas, conforme ilustrado na Figura 2. Tais etapas serão melhor descritas nas próximas seções e exemplificadas com as tecnologias selecionadas para a oficina.

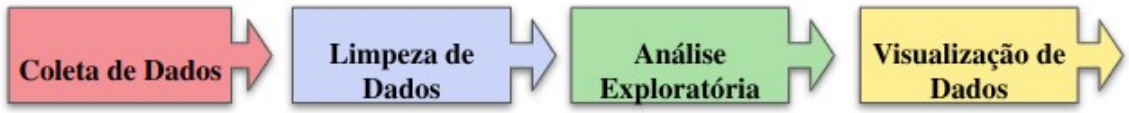

Figura 2 - Principais etapas da análise de dados.
Fonte: Elaborado pelos autores (2024).

4.1. Coleta de dados

Os dados são obtidos de diferentes fontes e podem estar estruturados ou não. Com o advento da Internet e evolução das tecnologias de comunicação, é cada vez mais comum as fontes de dados serem digitais e automatizadas de tal forma que cada vez menos processos manuais de coleta de dados são necessários. Os dados estruturados são organizados em um formato predefinido, como tabelas em bancos de dados. Por outro lado, dados não estruturados não possuem uma organização predefinida e podem incluir, por exemplo, textos, imagens, vídeos e outros formatos.

Nesta etapa, é essencial preparar o ambiente de trabalho com as bibliotecas necessárias instaladas e configuradas, além de carregar o *dataset* para análise. A partir disso, o analista de dados pode iniciar uma exploração preliminar do conjunto de dados. Os passos iniciais para essa análise incluem:

- Importação da biblioteca Pandas no projeto;
- Leitura do conjunto de dados CSV que representa o *dataset* usando a biblioteca Pandas. Se necessário conversão do dataset em formato apropriado;
- Renomeação das colunas do *DataFrame* e execução de funções descritivas para sumarizar informações;
- Diferenciar entre estruturas de dados como Séries e *DataFrames*; e
- Melhorar a visualização dos dados, gerando os primeiros gráficos para observar padrões e tendências.

4.2. Limpeza de dados

A etapa de limpeza de dados tem como objetivo normalizar os dados para o formato adequado para a análise. Nesta etapa os dados são limpos, duplicatas podem ser removidas, dados podem ser transformados ou normalizados de tal forma a melhorar a qualidade e a integridade. As tarefas comuns nesta etapa envolvem:

- Remoção de duplicatas usando a funcionalidade *drop_duplicates()* da bilioteca Pandas;
- Tratamento de valores ausentes por meio de imputação ou remoção;
- Correção de erros nos dados como tipografia e valores ausentes;
- Normalização dos dados numéricos ou categorização;
- Conversão de tipos de dados para facilitar o processamento e interpretação dos dados;
- Criação de variáveis derivadas a partir de preexistente com o intuito de melhorar a análise; e

- Filtragem dos dados.

4.3. Análise exploratória

A análise exploratória tem como objetivo identificar padrões, relações e possíveis tendências nos dados que possam fundamentar decisões e gerar conhecimento. Após a análise inicial realizada nas etapas de coleta e limpeza dos dados, é importante proceder com uma exploração mais profunda. A seguir, apresentamos algumas tarefas que podem ser realizadas para uma análise mais detalhada:

- Utilizar o método *query()* para filtrar linhas de um *DataFrame* conforme uma expressão definida;
- Agrupar os dados com base em uma coluna usando o método *groupby();*
- Extrair informações ao filtrar uma única coluna;
- Ajustar os intervalos (ou "*bins*") de um histograma para observar alterações no seu comportamento;
- Realizar a visualização dos dados usando gráficos de dispersão, histogramas, *boxplots* e barras;
- Realizar a análise estatística descritiva, como medidas de tendência central e dispersão;
- Avaliar a relação entre variáveis numéricas e identificar colinearidade por intermédio de matrizes de correlação;
- Identificação de *outliers*; e
- Analisar as tendências e padrões nas séries temporais e agrupamento dos dados.

Em um *DataFrame,* cada coluna representa um tipo específico de variável, que pode ser classificada como categórica ou quantitativa. A capacidade de distinguir e utilizar corretamente cada tipo de variável é fundamental para uma análise de dados eficaz. Portanto, os seguintes conhecimentos são essenciais:

- Identificar o tipo de variável observando seu conteúdo e estrutura;
- Distinguir uma variável categórica nominal de uma variável categórica ordinal; e
- Entender o conceito de variável quantitativa contínua e discreta e visualizar como elas diferem em relação aos intervalos de análise.

4.4. Visualização dos dados

Por fim, a etapa de visualização dos dados consiste na transformação dos dados complexos em visualizações que suportam a interpretação dos resultados, com intuito de comunicar de forma clara e eficaz o conhecimento e *insights* obtidos, onde as ferramentas visuais são de extrema importância para que informações sejam passadas corretamente, para isso, as seguintes tarefas precisam ser realizadas:

- Escolha do tipo de gráfico mais apropriados para o tipo de dado;
- Criação dos gráficos usando às funcionalidades das bibliotecas Seaborn e Matplotlib; e
- Personalização dos gráficos e visualização de séries temporais.

Considerando que a visualização gráfica desempenha um papel fundamental na análise de dados, é essencial realizar formatações nos gráficos para garantir que as informações sejam interpretadas de maneira clara e eficaz por aqueles que os visualizam.

- Ajustar as escalas de um gráfico para otimizar o espaço e aprimorar a visualização dos dados;
- Organizar os dados em um gráfico para conferir uma representação mais coerente;
- Modificar as cores e tonalidades de um gráfico para destacar elementos e melhorar sua estética; e
- Exibir um gráfico de colunas com valores absolutos e relativos.

5. Considerações Finais

A era da transformação digital trouxe consigo uma explosão de dados que têm o potencial de gerar *insights* valiosos e impactar diversas áreas da sociedade. A importância da análise de dados se evidencia nas aplicações práticas em setores como educação e saúde, permitindo a tomada de decisões rápidas e informadas que impactaram diretamente a vida de milhões de pessoas.

A oficina proposta neste documento não apenas introduz os participantes aos conceitos e técnicas fundamentais da análise de dados, mas também os capacita a aplicar essas habilidades em contextos reais, utilizando *datasets* amplamente reconhecidos. Com o aprendizado de ferramentas como Pandas, Matplotlib e Seaborn, os participantes possuirão conhecimentos básicos para enfrentar os desafios do mundo contemporâneo, onde a capacidade de interpretar dados se torna uma competência cada vez mais valorizada.

Por fim, espera-se que os conhecimentos adquiridos ao longo deste curso inspirem os participantes a aprofundar suas habilidades em análise de dados, contribuindo para suas trajetórias profissionais e acadêmicas permitindo que se tornem agentes de mudança em suas respectivas áreas de atuação.

Referências

CAVANILLAS, José María; CURRY, Edward; WAHLSTER, Wolfgang. *New horizons for a data-driven economy:* a roadmap for usage and exploitation of big data in Europe. 1. ed. Berlin: Springer, 2016.

DATABRICKS, Jupyter Notebook, Disponível em: https://www.databricks.com/br/glossary/jupyter-notebook. Acesso em: 30, out. 2024.

GOOGLE COLAB, Google Colaboratory, Disponível em: https://colab.google/. Acesso em 30, out. 2024.

GROUPLENS. MovieLens, Disponível em: https://files.grouplens.org/datasets/movielens/ml-latest-small.zip. Acesso em: 30 out. 2024

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. ed. New York: Springer, 2009.

KEAGLE, TMDB 500 Movie Dataset, Disponível em: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv. Acesso em: 30 out. 2024.

LANEY, Doug. 3D Data Management: Controlling Data Volume, Variety, and Velocity. Application Delivery Strategies, 2001.

PROVOST, Foster; FAWCETT, Tom. Data science for business: what you need to know about data mining and data-analytic thinking. 1. ed. Sebastopol, CA: O'Reilly Media, 2013.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews) 40.6 (2010): 601-618.

SIMPLILEARN. Data Science vs Data Analytics vs Machine Learning: What's the Difference? Disponível em: https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article. Acesso em: 30 out. 2024.

ZHANG, Jin; WOLFRAM, Dietmar; MA, Feicheng. The impact of big data on research methods in information science. Data and Information Management, v. 7, n. 2, 2023.