Análise Exploratória de Dados com Pandas.

Extraindo os principais insights ...

EDA

com Pandas

O objetivo da Análise Exploratória de Dados (EDA – Exploratory Data Analysis) é dar uma boa espiada nos dados antes de começar a fazer coisas mais complicadas. É como o investigador curioso que olha primeiro para entender do que se trata. A análise ajuda a descobrir segredos escondidos nos números, padrões estranhos e até erros, para que possamos tomar decisões mais inteligentes e contar histórias mais interessantes com nossos dados. É como a primeira pista em um quebra-cabeça gigante de informações.

rocketseat

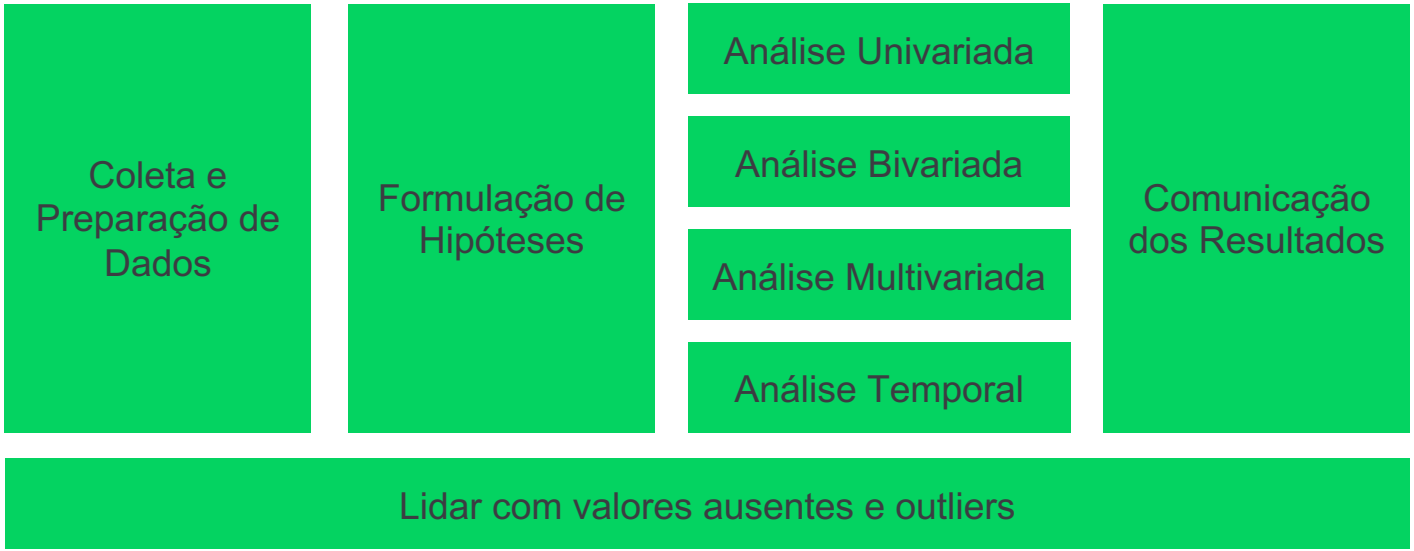
EDA

Com Pandas

01. O que é EDA?
02. O que é a biblioteca Pandas?
03. Coleta e Preparação de Dados
04. Lidando com dados ausentes
05. Formulando hipóteses
06. Análise Univariada
07. Análise Bivariada
08. Lidando com Outliers
09. Automatizando EDA

O que é EDA ?

rocketseat



EDA é um processo sistemático usado em projetos de ciência de dados para entender e resumir as características fundamentais de um conjunto de dados.

O que é a biblioteca Pandas

O Pandas é uma biblioteca de Python amplamente usada para análise de dados. Sua principal vantagem é sua capacidade de manipular, limpar e analisar dados de forma eficiente. Ele fornece estruturas de dados flexíveis, como Series e DataFrames, que permitem organizar dados em tabelas, realizar operações complexas, como filtros e agregações, e facilitar a visualização dos resultados. O Pandas também é compatível com várias fontes de dados, como arquivos CSV, Excel e bancos de dados, tornando-o essencial para cientistas de dados e analistas que desejam explorar e extrair insights de dados de maneira eficaz e intuitiva.

O contexto do nosso problema

O contexto que iremos analisar neste módulo de Análise Exploratória é do segmento de telecom. A empresa possui três conjuntos de dados de clientes e serviços, com uma variável que determina se o cliente abandonou (churn) ou não a empresa de telecom.

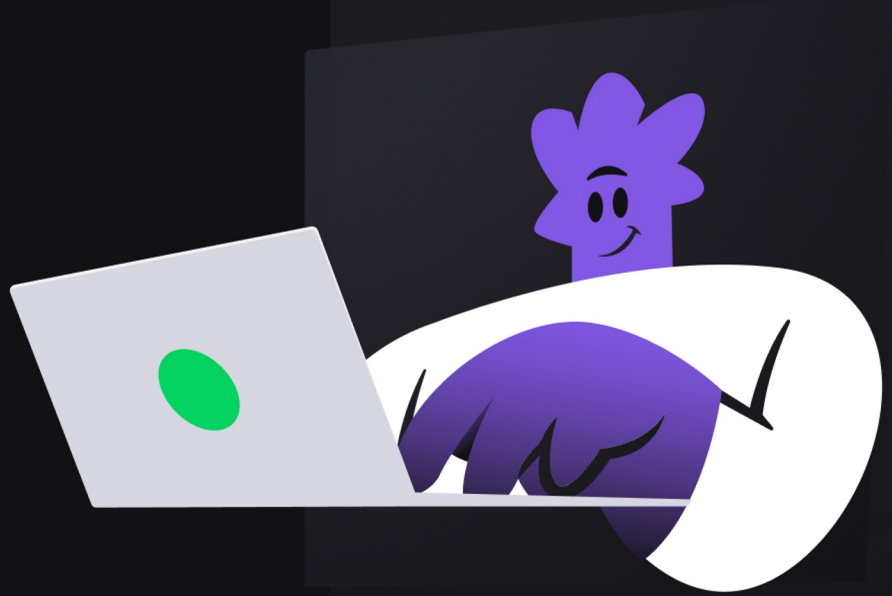
O intuito é que com estes conjuntos de dados e com base em algumas hipóteses que iremos formular e que serão respondidas pelo EDA, possa obter alguns insights iniciais para a construção de inteligência artificial que possa “prever” o abandono de clientes ainda ativos.

rocketseat

Coleta e Preparação de Dados

01. Leitura de arquivos
02. Transformação de tipos de dados
03. Renomear colunas
04. Unificação de conjuntos de dados

Let’s Go, Let’s Go



Code time ...

rocketseat

Lidar com dados ausentes

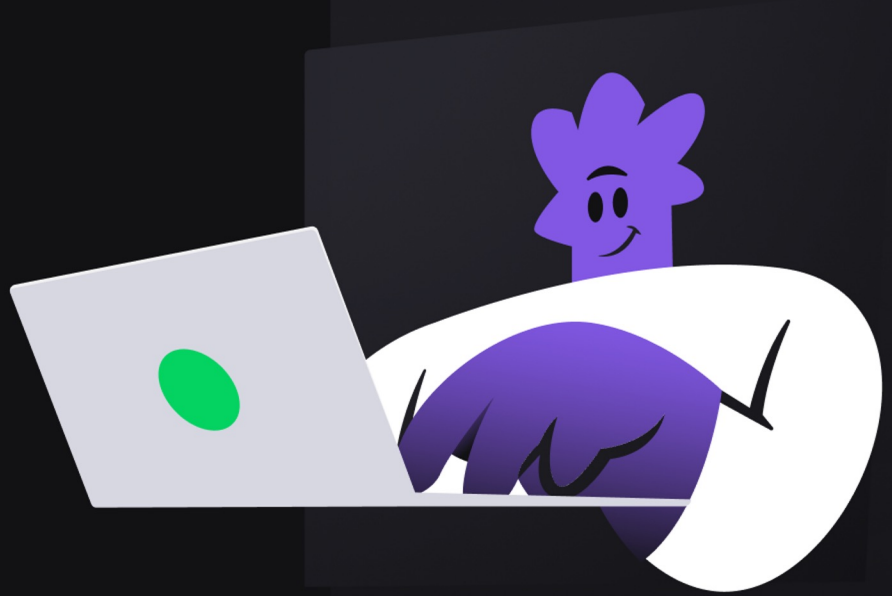
01. Tipos de dados ausentes
02. Detecção de dados ausentes
03. Remoção de dados ausentes
04. Imputação de dados ausentes

Tipos de dados ausentes

rocketseat

Dados faltantes completamente ao acaso (MCAR - Missing Completely at Random)	Dados faltantes ao acaso (MAR - Missing at Random)	Dados faltantes não ao acaso (MNAR - Missing not at Random)
O fato de que um certo valor está faltando não tem nada a ver com seu valor hipotético e com os valores de outras variáveis.	Faltar dados aleatoriamente significa que a propensão para um ponto de dados estar ausente não está relacionada aos dados ausentes, mas está relacionada a alguns dos dados observados.	Duas razões possíveis são que o valor ausente depende do valor hipotético ou o valor ausente depende do valor de alguma outra variável.

Let’s Go, Let’s Go



Code time ...

Formulando hipóteses

rocketseat

Use a Intuição	Seja Específico	Testabilidade	Considere Relações
Comece com suas suspeitas iniciais com base no conhecimento do domínio. Pergunte a si mesmo o que você espera encontrar nos dados.	Suas hipóteses devem ser claras e específicas. Evite afirmações vagas, como "os dados têm alguma tendência". Em vez disso, seja concreto, como "o aumento nas vendas está relacionado ao lançamento de um novo produto".	Certifique-se de que suas hipóteses possam ser testadas com os dados disponíveis. Você deve ser capaz de encontrar evidências nos dados que confirmem ou refutem a hipótese.	Pense em como diferentes variáveis podem estar relacionadas. Por exemplo, "a idade dos clientes afeta a taxa de churn?" ou "a localização geográfica influencia as preferências de compra?".

Formulando hipóteses

rocketseat

Utilize Visualizações	Seja Aberto a Surpresas	Evite Hipóteses de Causa e Efeito Prematuras	Refine e Atualize
A EDA frequentemente envolve gráficos e visualizações. Use-os para explorar seus dados e validar suas hipóteses. Por exemplo, crie gráficos de dispersão para verificar relações entre variáveis.	Esteja preparado para encontrar resultados inesperados. Às vezes, as melhores descobertas ocorrem quando as hipóteses iniciais são desafiadas.	Evite assumir causalidade imediatamente. Por exemplo, "A promoção de um produto causou um aumento nas vendas". Primeiro, explore a correlação e, em seguida, avalie a causalidade.	À medida que você realiza a EDA e coleta mais informações, refine e atualize suas hipóteses conforme necessário. Este é um processo iterativo.

Formulação de Hipóteses

01. A faixa etária do cliente tem uma forte associação com o churn
02. Um cliente que com menos de 6 meses de contrato ativo é mais propenso ao Churn
03. Cliente com contratos mensais são mais propensos ao Churn

Análise Univariada

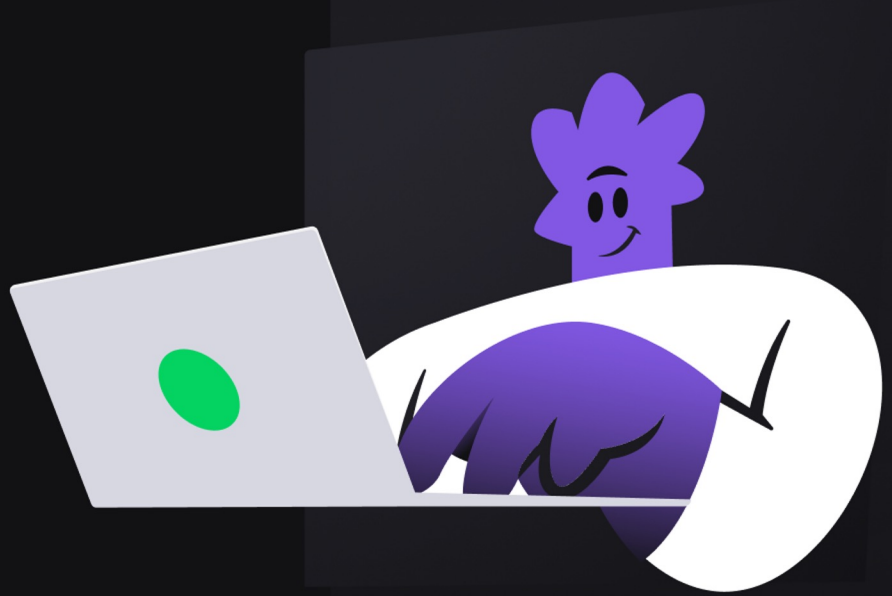
A análise univariada é uma abordagem estatística que se concentra na análise de uma única variável em um conjunto de dados. Ela visa compreender as características individuais dessa variável, examinando sua distribuição, medidas resumo (como média e mediana), variabilidade e a presença de valores atípicos (outliers). Isso ajuda a obter uma visão detalhada das características de uma variável específica, antes de explorar relações com outras variáveis (análise bivariada ou multivariada) durante a análise de dados.

rocketseat

Análise Univariada

01. Distribuição
02. Medidas de Posição
03. Medidas de Dispersão

Let’s Go, Let’s Go



Code time ...

Análise Bivariada

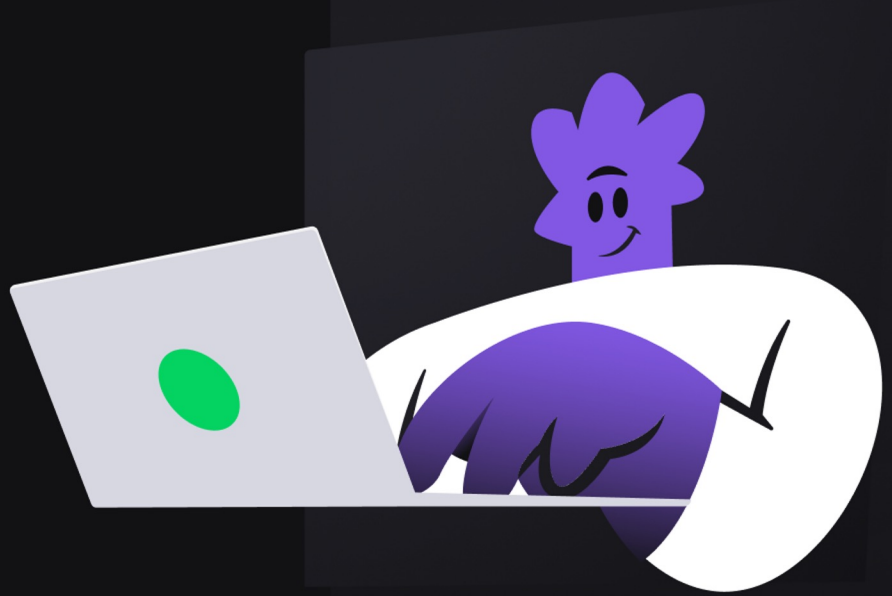
A análise bivariada é uma técnica estatística que se concentra na relação entre duas variáveis em um conjunto de dados. Ela busca entender como uma variável está relacionada à outra, frequentemente usando gráficos, tabelas cruzadas e cálculos de correlação. Isso ajuda a identificar padrões, associações e dependências entre as duas variáveis, fornecendo insights sobre como elas interagem, o que é crucial na análise de dados e na tomada de decisões informadas.

rocketseat

Análise Bivariada

01. Análise de Correlação
02. Tabelas Cruzadas

Let’s Go, Let’s Go



Code time ...

Outliers (Valores atípicos)

Um outlier é um dado que é muito diferente dos outros dados em um conjunto de dados. É como um ponto fora da curva.

Por exemplo, imagine que você tem um conjunto de dados que registra a altura de 100 pessoas. A média das alturas é de 1,70 metros. Um outlier seria uma pessoa que tem 2,50 metros de altura. Essa pessoa é muito mais alta do que as outras, então ela é considerada um outlier.

Outliers podem ser causados por vários fatores, como erros de medição, dados incompletos ou eventos aleatórios. Eles podem afetar os resultados de uma análise de dados, então é importante identificá-los e lidar com eles de forma adequada.

Lidando com Outliers

rocketseat

Identificação e Documentação	Remoção	Transformação de Dados	Binning (Agrupamento)
Identifique os outliers em seus dados e documente-os. Compreender a natureza dos outliers é fundamental.	Em alguns casos, você pode optar por remover os outliers do conjunto de dados, desde que isso seja justificável e não distorça a análise.	Aplicar transformações matemáticas aos dados, como logaritmo ou raiz quadrada, pode reduzir o impacto dos outliers.	Dividir os dados em intervalos (bins) pode ajudar a reduzir a influência de outliers ao transformar a variável em categórica.

Lidando com Outliers

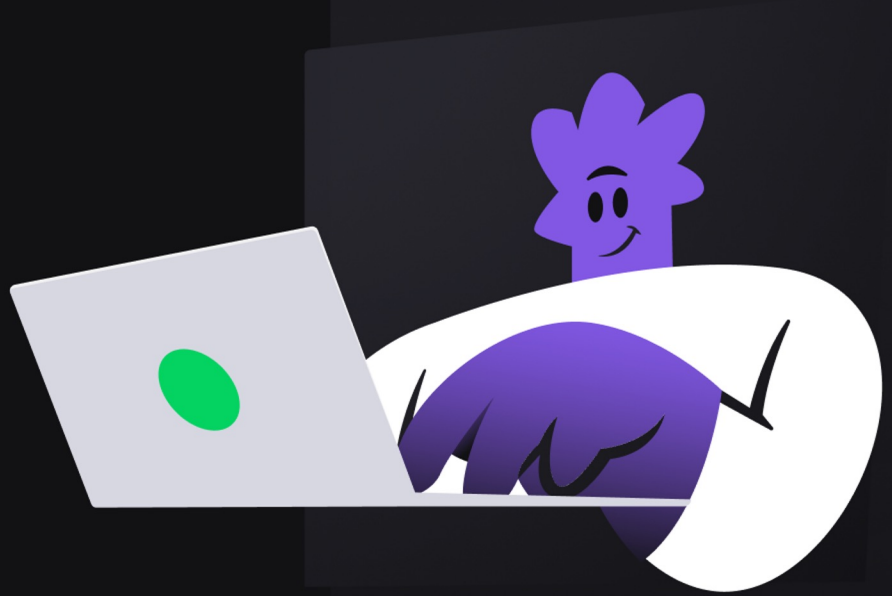
rocketseat

Substituição	Modelagem Robusta	Análise Separada	Entender a Causa
Substituir outliers por valores mais representativos, como a mediana, a média truncada ou valores interpolados.	Usar algoritmos de modelagem mais robustos a outliers, como regressão robusta ou métodos de aprendizado de máquina robustos.	Realizar análises separadas para dados com e sem outliers para avaliar o impacto deles nas conclusões.	Se possível, investigar a causa dos outliers para determinar se são erros de medição ou representam informações legítimas.

Lidando com Outliers

01. Detecção de Outliers

Let’s Go, Let’s Go



Code time ...

Automatizando EDA

A automatização da análise exploratória de dados (EDA) oferece uma série de vantagens, incluindo:

Aumento da velocidade e da eficiência: A EDA automatizada pode ser executada muito mais rapidamente do que a EDA manual. Isso pode ser importante para conjuntos de dados grandes ou complexos.

Redução da subjetividade: A EDA automatizada é menos propensa a erros ou vieses humanos. Isso pode levar a análises mais precisas e confiáveis.

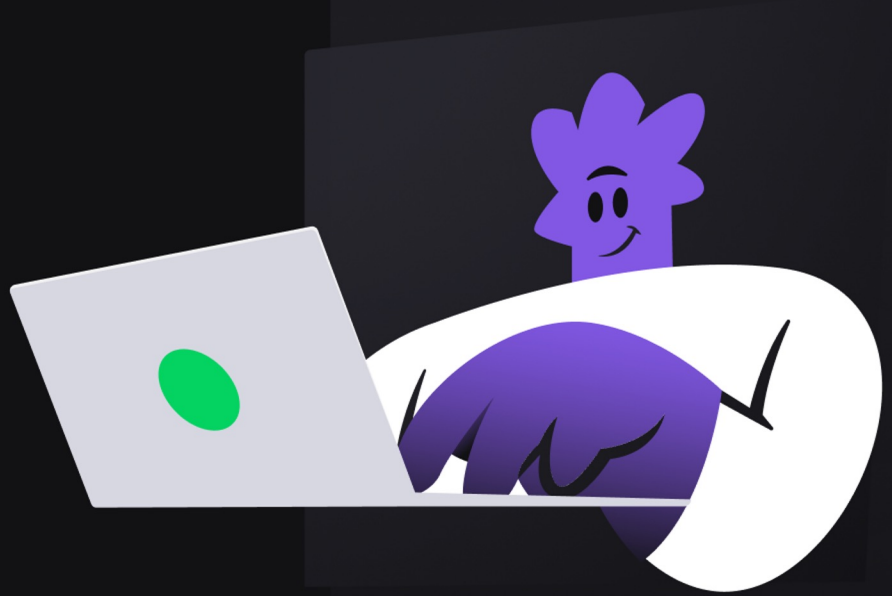
Melhor compreensão dos dados: A EDA automatizada pode identificar padrões e tendências que podem não ser óbvios para os analistas humanos. Isso pode ajudar a obter uma melhor compreensão dos dados e a tomar melhores decisões.

rocketseat

Automatizando EDA

01. Sweetviz

Let’s Go, Let’s Go



Code time ...

rocketseat

Boosting People.

rocketseat.com.br