| CURSO: TECNOLOGIA EM CIÊNCIA DE DADOS |
| --- |
| POLO DE APOIO PRESENCIAL: |
| SEMESTRE: |
| COMPONENTE CURRICULAR / TEMA: PROJETO APLICADO I – 1ª ETAPA |
| NOME COMPLETO DO ALUNO: |
| TIA: |
| NOME DO PROFESSOR: FELIPE ALBINO DOS SANTOS |

* **Título do trabalho:**

Compreendendo a Dinâmica do Mercado na Amazon.

* **Membros do grupo:**

Nome: Alex Júnior Moura da Silva

RA: 10415336

E-mail: [10415336@mackenzista.com.br](mailto:10415336@mackenzista.com.br)

Link GitHub: [mack-ciencia-de-dados/projeto-aplicado-1-2024 (github.com)](https://github.com/mack-ciencia-de-dados/projeto-aplicado-1-2024)

Nome: Felipe Akira Fukue

RA: 10424046

E-mail: 10424046@mackenzista.com.br

Link GitHub: [mack-ciencia-de-dados/projeto-aplicado-1-2024 (github.com)](https://github.com/mack-ciencia-de-dados/projeto-aplicado-1-2024)

Nome: Wanessa Carvalho dos Santos

RA: 10416116

E-mail: @mackenzista.com.br

Link GitHub: [mack-ciencia-de-dados/projeto-aplicado-1-2024 (github.com)](https://github.com/mack-ciencia-de-dados/projeto-aplicado-1-2024)

Nome: Gustavo Guarizzo Esteves

RA: 10424491

E-mail: 10424491@mackenzista.com.br

Link GitHub: [mack-ciencia-de-dados/projeto-aplicado-1-2024 (github.com)](https://github.com/mack-ciencia-de-dados/projeto-aplicado-1-2024)

* **Referências de Aquisição do Dataset:** dataset adquirido pela plataforma Kaggle. Este dataset contém dados de mais de mil classificações e análises de produtos da Amazon de acordo com seus detalhes listados no site oficial da empresa. O dataset pode ser encontrado pelo link: https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset

O dataset está sob a licença: CC BY-NC-SA 4.0 (<https://creativecommons.org/licenses/by-nc-sa/4.0/>)

Período de coleta: Pelo perfil da base de dados, não tem uma referência do período da coleta de dados, porém os mesmos foram coletados a um ano atrás (2023)

* **Descrição da Origem:** Os dados utilizados neste projeto foram adquiridos através da plataforma Kaggle, uma comunidade online que oferece conjuntos de dados diversificados e desafios relacionados à ciência de dados. O Kaggle se destaca como uma plataforma integral que não apenas oferece desafios de ciência de dados, mas também fornece recursos educacionais, ferramentas de colaboração e uma comunidade engajada para impulsionar o aprendizado e avanço na área de ciência de dados.
* **Descrição do dataset:**

A base de dados possui as seguintes colunas com suas respectivas descrições:

* product\_id - ID do Produto
* product\_name - Nome do Produto
* category - Categoria do Produto
* discounted\_price - Preço com Desconto do Produto
* actual\_price - Preço Real do Produto
* discount\_percentage - Porcentagem de Desconto do Produto
* rating - Classificação do Produto
* rating\_count - Número de pessoas que votaram na classificação da Amazon
* about\_product - Descrição sobre o Produto
* user\_id - ID do usuário que escreveu a avaliação para o Produto
* user\_name - Nome do usuário que escreveu a avaliação para o Produto
* review\_id - ID da avaliação do usuário
* review\_title - Resenha curta
* review\_content - Resenha longa
* img\_link - Link da Imagem do Produto
* product\_link - Link do Site Oficial do Produto

Nele foram registrados engajamentos do usuário com os produtos da Amazon.

* **Contexto do Estudo:** O objetivo central deste projeto é explorar e entender as relações entre as variáveis-chave dos produtos listados na Amazon, uma das maiores plataformas de comércio eletrônico do mundo. Especificamente, a análise se concentrará em discernir as interconexões entre as avaliações dos clientes (ratings), os preços, os descontos aplicados e as categorias dos produtos. O estudo procura identificar padrões e correlações significativas entre esses fatores, visando compreender melhor como eles se influenciam mutuamente. Esse entendimento pode lançar luz sobre a dinâmica do mercado e as preferências dos consumidores, mesmo sem fazer suposições diretas sobre o comportamento de compra.
* **Objetivos e metas:** O objetivo geral do projeto é realizar uma análise exploratória de dados na base adquirida, aplicando técnicas avançadas de dados e analytics para desvendar as relações complexas entre as variáveis-chave dos produtos da Amazon, incluindo avaliações de clientes, preços, descontos e categorias de produtos.

1. ***Exploração do Dataset:*** Realizar uma exploração inicial do conjunto de dados da Amazon para entender sua estrutura, tipos de variáveis e possíveis desafios.
2. ***Limpeza e Preparação dos Dados:***  Limpar e preparar os dados para a análise, tratando valores ausentes, removendo duplicatas e ajustando formatos de datas.
3. ***Análise Descritiva Inicial:*** Aplicar técnicas de pensamento computacional para realizar análises descritivas iniciais, como médias, desvios padrão e quantis.
4. ***Visualizações Significativas:*** Criar visualizações exploratórias, com gráficos e análises claras que auxiliem na comunicação eficaz dos insights obtidos.
5. ***Relatório de conclusões:*** Criação de um relatório claro e bem documentado, descrevendo as análises realizadas, os métodos empregados e os resultados obtidos.

* **Cronograma de atividades:** Este projeto será dividido em quatro etapas distintas, cada um com seus objetivos específicos e atividades:

1. ***Kick-off:***

* Formação da equipe do projeto;
* Definição de metas e objetivos detalhados;
* Coleta e revisão inicial dos dados obtidos.

1. ***Definição do Produto Analítico:***

* Preparação e limpeza dos dados;
* Desenvolvimento de métricas-chave para avaliação;
* Identificação de abordagens analíticas e ferramentas necessárias.

1. ***Data Storytelling:***

* Análise exploratória dos dados;
* Identificação de tendências e padrões;
* Criação de narrativas visuais para comunicar os resultados.

1. ***Apresentação dos Resultados:***

* Preparação do relatório final;
* Apresentação dos resultados.

1. ***Coleta de Dados:*** Garantir que os dados sejam adquiridos e armazenados de maneira eficiente e segura.
2. ***Formulação de Problemas:*** Identificar questões e desafios relacionados às vendas da Amazon que podem ser abordados através de análises computacionais. Por exemplo, como entender padrões sazonais, correlacionar eventos com variações nas vendas, ou prever tendências futuras.
3. ***Análise de Dados:*** Aplicar abordagens de análise de dados, como modelagem preditiva e análise de séries temporais, para extrair insights dos dados da Amazon.
4. ***Pensamento Lógico:*** Aplicar pensamento lógico para interpretar os resultados das análises e derivar conclusões significativas.
5. ***Resolução de Problemas de Forma Iterativa:*** Adotar uma abordagem iterativa para resolver problemas, ajustando análises com base em descobertas e feedbacks ao longo do projeto.
6. ***Comunicação dos Resultados:*** Utilizar visualizações de dados e técnicas de data storytelling para comunicar os resultados de maneira acessível.
7. ***Tomada de Decisões:*** Capacitar os tomadores de decisões com as conclusões alcançadas durante a análise e fornecer recomendações práticas para estratégias futuras.

* **Análise Exploratória:**

**1. Objetivo de Estudo**

**Apresentação dos Dados da Empresa:**

* + **Nome da Empresa:** Amazon
  + **Missão/Visão/Valores:** A Amazon se orienta por 4 pilares: obsessão pelo cliente, paixão por invenções, compromisso com excelência operacional e visão de longo prazo.
  + **Segmento de Atuação:** A Amazon atua em vários segmentos de mercado, incluindo e-commerce, cloud computing (AWS), inteligência artificial, streaming de conteúdo digital e produção de eletrônicos de consumo.
  + **Market Share/Posicionamento no Mercado:** A Amazon é líder global no segmento de e-commerce e detém uma parcela significativa do mercado de serviços em nuvem com a AWS. Seu posicionamento é de uma empresa inovadora que continua a expandir sua influência em novos setores de mercado.
  + **Número de Colaboradores: A** Amazon é uma das maiores empregadoras do mundo, com centenas de milhares de funcionários em vários países.
  + **Iniciativas na Área de Data Science:** A Amazon investe fortemente em data science e machine learning, utilizando essas tecnologias para melhorar a experiência do cliente, otimizar logística, prever tendências de mercado e inovar em personalização e recomendações de produtos. AWS também oferece uma gama de serviços e ferramentas para data scientists e engenheiros de machine learning.

**2. Perguntas a responder**

Neste processo, procuramos responder às seguintes perguntas:

* + - * **Qual é o número de exemplares (linhas) e dimensões (colunas)?**

O dataset possui 1465 linhas e 16 colunas.

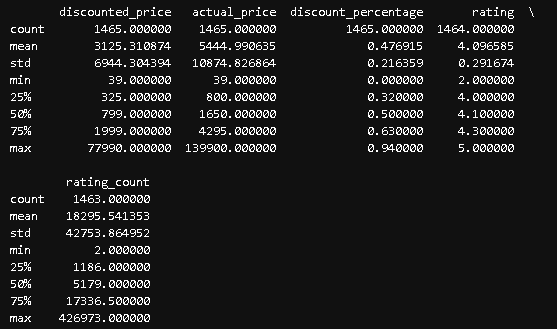
* + - * **Quais são os tipos de dados presentes?**

Com base na natureza estatística dos dados, aqui está como poderíamos categorizar as colunas do dataset amazon.csv:

* + **product\_id:** Variável nominal, pois é um identificador que categoriza cada produto de forma única.
  + **product\_name:** Também é uma variável nominal, usada para categorizar produtos por nomes.
  + category: Variável nominal que agrupa produtos em diferentes categorias.
  + **discounted\_price e actual\_price:** Variáveis racionais (contínuas), pois representam quantidades que podem assumir uma gama infinita de valores dentro de um intervalo.
  + **discount\_percentage:** Uma variável racional (contínua) que indica a proporção do desconto em relação ao preço total.
  + **rating:** Uma variável discreta, e representa um ranking em uma escala com ordem (1-5).
  + **rating\_count:** Uma variável discreta (quantitativa) que conta o número de avaliações.
  + **about\_product, review\_title e review\_content:** Variáveis textuais, podem ser consideradas nominais para a categorização do conteúdo textual.
  + **user\_id e user\_name:** Variáveis nominais devido à sua natureza de identificação.
  + **review\_id:** Variável nominal que serve como um código único para cada avaliação.
  + **img\_link e product\_link:** Variáveis nominais que fornecem referências a URLs.

Ao converter para os tipos estatísticos apropriados, podemos aplicar técnicas de análise descritiva, inferencial e preditiva mais precisas, cada uma alinhada com as características dos dados em questão.

* + - * **Quais são as medidas de posição e dispersão?**

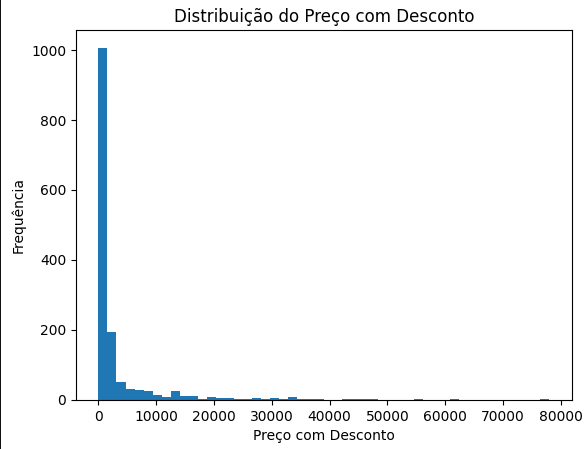


Esta imagem contém as medidas de posição e dispersão.

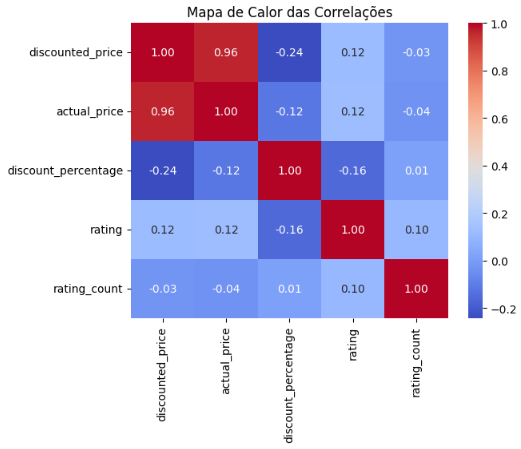
* + - * **Como é a distribuição e frequência dos dados?**

Os 5 mais frequentes:

* + Computers&Accessories|Accessories&Peripherals|Cables&Accessories|Cables|USBCables
  + Electronics|WearableTechnology|SmartWatches
  + Electronics|Mobiles&Accessories|Smartphones&BasicMobiles|Smartphones
  + Electronics|HomeTheater,TV&Video|Televisions|SmartTelevisions
  + Electronics|Headphones,Earbuds&Accessories|Headphones|In-Ear



* + - * **Existem correlações significativas entre as variáveis?**



Observando o gráfico, entendemos que essa correlação alta entre o preço com desconto e o preço real, de 0.96, mostra que quanto mais caro é o produto, mais alto é o desconto em valor absoluto. É aquela história, quanto mais você gasta, mais você "economiza", né?

Já a relação do percentual de desconto com o preço: é negativa. Ou seja, não espere um descontão percentual nos itens mais caros – a loja não está dando aquele abate tão grande em porcentagem quanto nos produtos mais em conta.

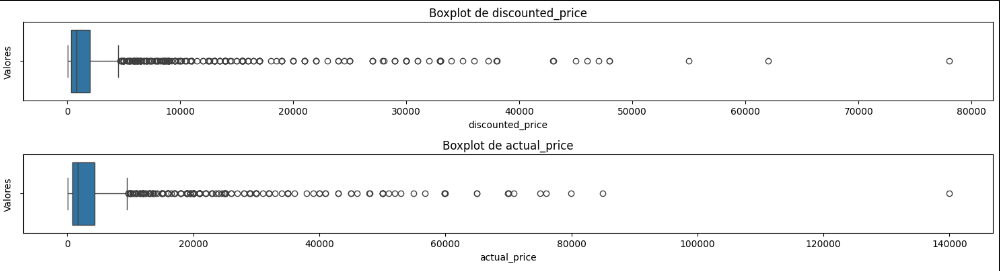
Agora, as notas dos produtos e a quantidade de avaliações não estão dando uma de melhores amigos com o preço ou o desconto. Não é porque o produto é mais caro ou tem um super desconto que ele vai ter notas altas ou um montão de comentários.

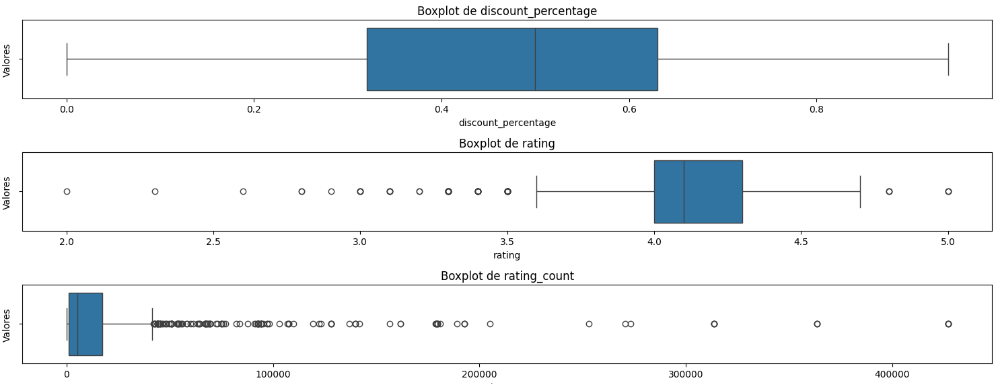
E tem mais um detalhe: tem uma leve tendência de produtos mais caros levarem notas um pouco melhores, o que pode ser um indício de que preço mais alto pode ser sinal de qualidade, ou pelo menos de percepção de qualidade.

* + - * **Há valores perdidos ou incorretos?**

Fazendo o devido tratamento e consulta foi descoberto que havia valores perdidos, mas eram em quantidade mínima nas colunas “rating” e “rating\_count”.

* + - * **Existem anomalias ou outliers?**





Nos preços com desconto (discounted\_price) e nos preços reais (actual\_price), tem uma quantidade bem grande de outliers para cima. Isso pode indicar que a maioria dos produtos tem um preço mais em conta, mas tem uns poucos que são bem mais caros. Talvez sejam produtos de categoria premium ou algo assim.

Já na porcentagem de desconto (discount\_percentage), o boxplot está bem compacto e não mostra outliers. Parece que os descontos são bem consistentes, mas tem uns poucos produtos com desconto quase zero, que estão ali no limite superior do bigode do boxplot.

O rating tá mais ou menos como a gente esperaria para um e-commerce: a maioria dos produtos tem notas entre 3 e 5, com poucos produtos tendo notas superbaixas. Os outliers estão acima de 4.5, o que sugere que tem produtos que são muito bem avaliados pela galera.

Agora, em relação ao rating\_count, temos muitos outliers. Isso mostra que tem uns produtos que são muito mais avaliados que a média. Pode ser que esses produtos sejam muito populares ou estão há mais tempo no mercado.

* + - * StoryTelling: Esboço para apresentação

Título do trabalho:

Compreendendo a Dinâmica do Mercado na Amazon.

https://github.com/mouralex/projeto-aplicado-1/tree/master

• Nome da empresa; Amazon

Area do problema; área do problema parece estar relacionada ao comércio eletrônico, com um foco específico na plataforma da Amazon. Especificamente, o problema se concentra em entender as relações entre variáveis-chave dos produtos listados na Amazon, como avaliações de clientes, preços, descontos e categorias de produtos. Isso sugere uma área de interesse que abrange tanto a análise de dados quanto os princípios de marketing e comércio eletrônico. Com ênfase na compreensão das relações entre as variáveis mencionadas.

CONTEXTUALIZAÇÃO:

Descrição do Problema: a necessidade de explorar e compreender as relações entre variáveis-chave dos produtos listados na Amazon, uma importante plataforma de comércio eletrônico. Especificamente, o foco está em entender as interconexões entre as avaliações dos clientes, os preços, os descontos aplicados e as categorias dos produtos. O objetivo é identificar padrões e correlações significativas entre esses fatores para compreender melhor como eles se influenciam mutuamente. Essa compreensão pode oferecer insights sobre a dinâmica do mercado e as preferências dos consumidores, sem fazer suposições diretas sobre o comportamento de compra.

Entendimento do mercado: Compreender as relações entre as variáveis-chave dos produtos na Amazon pode oferecer insights valiosos sobre a dinâmica do mercado. Isso permite que os vendedores e as empresas ajustem suas estratégias de marketing, precificação e promoção para se adaptarem melhor às demandas dos consumidores.

Tomada de decisão informada: Ao identificar padrões e correlações significativas entre as avaliações dos clientes, preços, descontos e categorias de produtos, as empresas podem tomar decisões mais informadas. Isso inclui decisões relacionadas à seleção de produtos, definição de preços competitivos e planejamento de promoções.

Melhoria da experiência do cliente: Compreender como essas variáveis se influenciam mutuamente pode ajudar as empresas a melhorar a experiência do cliente. Por exemplo, ao ajustar os preços com base nas avaliações dos clientes ou oferecer descontos em categorias específicas de produtos com alta demanda, as empresas podem aumentar a satisfação do cliente e a fidelidade à marca.

Identificação de tendências emergentes: Analisar as interconexões entre as variáveis-chave dos produtos na Amazon pode ajudar a identificar tendências emergentes no mercado. Isso permite que as empresas se adaptem rapidamente às mudanças nas preferências dos consumidores e capitalizem oportunidades de mercado.

* + - * Proposta Analítica: explorar os dados disponíveis para compreender a dinâmica do mercado na Amazon, identificar insights valiosos e fornecer recomendações acionáveis para melhorar o desempenho e a competitividade dos produtos na plataforma. Envolvendo;

Exploração das Relações entre Variáveis: Utilizar técnicas estatísticas e de visualização de dados para compreender as relações entre as variáveis-chave dos produtos da Amazon, como avaliações de clientes, preços, descontos e categorias de produtos.

Identificação de Padrões e Correlações Significativas: Aplicar métodos estatísticos e algoritmos de aprendizado de máquina para identificar padrões e correlações significativas nos dados. Isso pode incluir análise de correlação, análise de regressão e técnicas de clustering para agrupar produtos com características semelhantes.

Análise de Sentimento das Avaliações dos Clientes: Utilizar técnicas de processamento de linguagem natural (PLN) para realizar uma análise de sentimento das avaliações dos clientes. Isso pode ajudar a entender melhor como as opiniões dos clientes impactam as vendas e a reputação dos produtos.

Segmentação de Mercado: Segmentar o mercado com base nas características dos produtos e dos clientes. Isso pode ajudar a identificar grupos de produtos que atendem a diferentes segmentos de clientes e adaptar estratégias de marketing e precificação para cada segmento.

Previsão de Tendências e Demanda: Desenvolver modelos de previsão para antecipar tendências de mercado e demanda por diferentes produtos. Isso pode ser útil para otimizar o estoque, gerenciar a cadeia de suprimentos e planejar estratégias de lançamento de novos produtos.

DADOS DISPONÍVEIS:

• Fonte de dados/ Descrição dos Dados:

* + **product\_id:** Variável nominal, pois é um identificador que categoriza cada produto de forma única.
  + **product\_name:** Também é uma variável nominal, usada para categorizar produtos por nomes.
  + category: Variável nominal que agrupa produtos em diferentes categorias.
  + **discounted\_price e actual\_price:** Variáveis racionais (contínuas), pois representam quantidades que podem assumir uma gama infinita de valores dentro de um intervalo.
  + **discount\_percentage:** Uma variável racional (contínua) que indica a proporção do desconto em relação ao preço total.
  + **rating:** Uma variável discreta, e representa um ranking em uma escala com ordem (1-5).
  + **rating\_count:** Uma variável discreta (quantitativa) que conta o número de avaliações.
  + **about\_product, review\_title e review\_content:** Variáveis textuais, podem ser consideradas nominais para a categorização do conteúdo textual.
  + **user\_id e user\_name:** Variáveis nominais devido à sua natureza de identificação.
  + **review\_id:** Variável nominal que serve como um código único para cada avaliação.
  + **img\_link e product\_link:** Variáveis nominais que fornecem referências a URLs.

**-**

* product\_id - ID do Produto
* product\_name - Nome do Produto
* category - Categoria do Produto
* discounted\_price - Preço com Desconto do Produto
* actual\_price - Preço Real do Produto
* discount\_percentage - Porcentagem de Desconto do Produto
* rating - Classificação do Produto
* rating\_count - Número de pessoas que votaram na classificação da Amazon
* about\_product - Descrição sobre o Produto
* user\_id - ID do usuário que escreveu a avaliação para o Produto
* user\_name - Nome do usuário que escreveu a avaliação para o Produto
* review\_id - ID da avaliação do usuário
* review\_title - Resenha curta
* review\_content - Resenha longa
* img\_link - Link da Imagem do Produto
* product\_link - Link do Site Oficial do Produto

ANÁLISE EXPLORATÓRIA

Metodologia de Análise Exploratória: Os dados utilizados neste projeto foram adquiridos através da plataforma Kaggle, uma comunidade online que oferece conjuntos de dados diversificados e desafios relacionados à ciência de dados. O Kaggle se destaca como uma plataforma integral que não apenas oferece desafios de ciência de dados, mas também fornece recursos educacionais, ferramentas de colaboração e uma comunidade engajada para impulsionar o aprendizado e avanço na área de ciência de dados.

RESULTADOS PRETENDIDOS

Objetivos do projeto: O objetivo do projeto é realizar uma análise exploratória de dados na base de dados adquirida, utilizando técnicas avançadas de dados e análise para desvendar as relações complexas entre as variáveis-chave dos produtos da Amazon. Essas variáveis incluem as avaliações dos clientes, os preços, os descontos e as categorias de produtos. Em essência, o objetivo é compreender melhor como essas variáveis se relacionam entre si, buscando identificar padrões e correlações significativas que possam oferecer insights valiosos para a compreensão da dinâmica do mercado e das preferências dos consumidores.

• Impacto Esperado: pode ser significativo em várias áreas:

Ao realizar uma análise profunda das relações entre as variáveis-chave dos produtos na Amazon, podemos oferecer uma compreensão mais completa da dinâmica do mercado na plataforma. Isso não apenas beneficia a Amazon, mas também os vendedores terceirizados que utilizam a plataforma para comercializar seus produtos. As descobertas e insights gerados podem ajudar a Amazon e seus vendedores parceiros a otimizar suas estratégias de negócios. Isso inclui ajustar preços, promoções e estratégias de marketing com base em padrões identificados nos dados.