

# Sobre a Amazon

---

- O site Amazon foi lançado no ano de 1995 por **Jeff Bezos** e hoje quase metade de todas as vendas on-line nos EUA é feita pela Amazon, tornando-a a empresa líder neste mercado além de ocupar o segundo lugar em termos de volume de mercadorias globais vendidas.

“A missão da Amazon é ser a empresa com mais **foco no cliente** do mundo todo e se esforçar para oferecer os menores preços possíveis de maneira que os clientes consigam encontrar e descobrir tudo o que desejem comprar.” - Jeff Bezos



# Antes da Inteligência Artificial

A Amazon utiliza inteligência artificial para impulsionar seus produtos, aumentar vendas, prevenir fraudes e melhorar a **experiência** do cliente. Isso inclui:

- Recomendações personalizadas;
- Chatbots e assistentes virtuais;
- Detecção de fraudes;
- Otimização logística;

Porém, antes que a IA possa se tornar uma solução que agregue valor, é necessário adquirir um amplo conhecimento por meio de **análises exploratórias**. As análises exploratórias envolvem a investigação e o entendimento dos dados disponíveis, identificando padrões, tendências e insights relevantes. Essa etapa é fundamental para garantir que os algoritmos de IA sejam alimentados com informações confiáveis e relevantes.



# O Python e o pré-processamento dos dados

A análise exploratória de dados requer linguagens de programação com ferramentas adequadas para o tratamento dos dados. Essas ferramentas são essenciais para realizar melhores visualizações, transformações, manipulações, agregações e cálculos que **enriquecem os dados**, garantindo a qualidade e preparando-os para uso nos algoritmos de IA. Elas também desempenham um papel fundamental na identificação **e tratamento** de anomalias e dados inconsistentes, garantindo uma análise abrangente e eficaz.

Na análise exploratória da Amazon, utilizamos o **Python** e suas bibliotecas:

- Pandas, os e glob – Para importação e junção dos arquivos CSV
- Pandas – Para verificar a dimensão dos dados
- Pandas – Para validar os tipos de dados
- Pandas, Numpy, e funções recursivas – Para tratar valores perdidos ou incorretos
- Pandas – Para verificar as medidas de posição e dispersão
- Seaborn, Pandas e Matplotlib – Para a criação do gráficos para visualização dos dados

	ratings	no_of_ratings	discount_price	actual_price
count	774717.000000	787341.000000	1.039503e+06	1.122810e+06
mean	3.839880	81.522315	3.348425e+02	2.566563e+02
std	0.753968	166.752112	3.007437e+02	3.709120e+02
min	1.000000	0.000000	1.000000e+00	0.000000e+00
25%	3.500000	3.000000	6.000000e+00	1.999000e+00
50%	3.900000	12.000000	2.990000e+02	4.999000e+00
75%	4.300000	65.588000	5.460000e+02	5.490000e+02
max	5.000000	999.000000	9.999900e+02	9.999900e+02

`df.describe()` para medidas de posição e dispersão

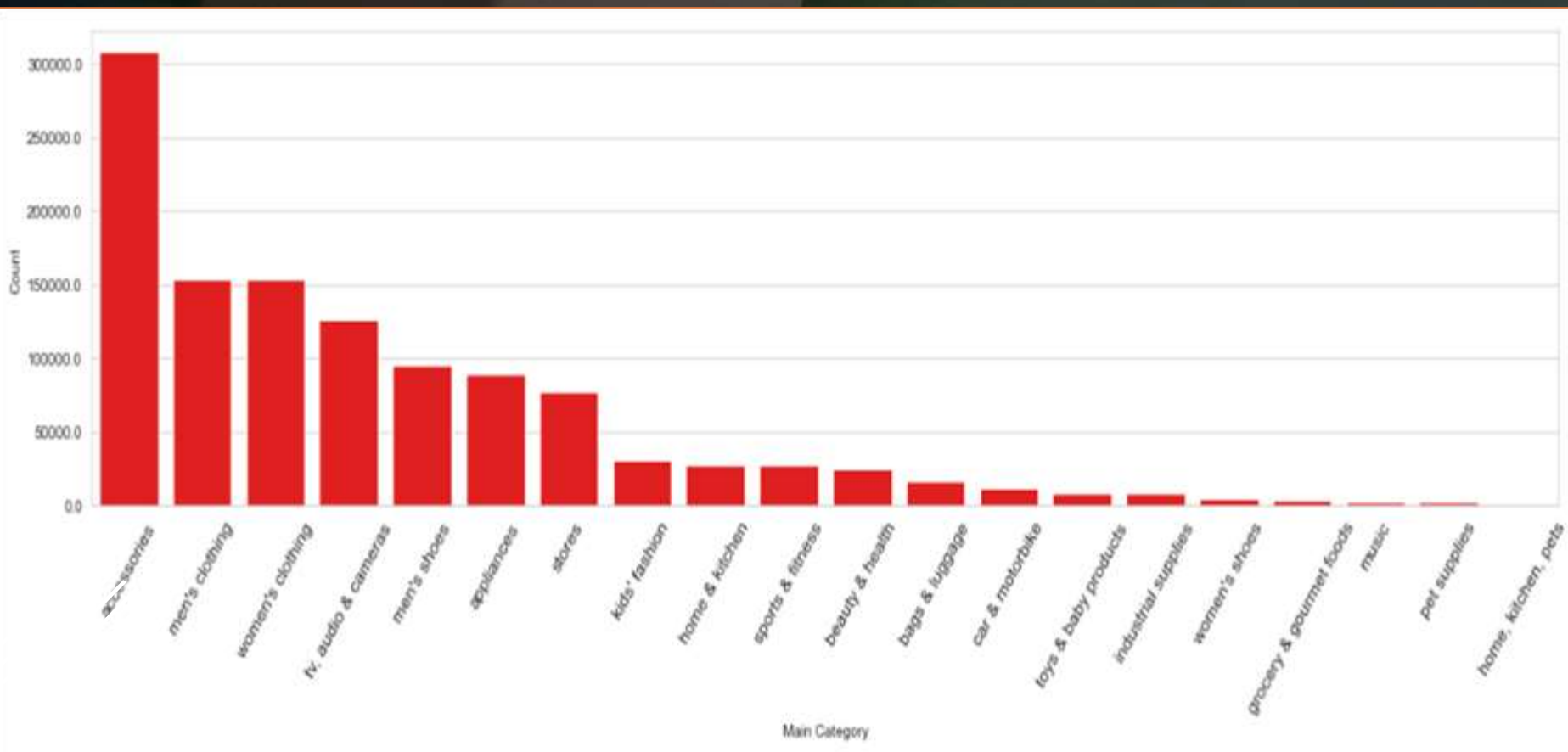


# Visualização dos resultados obtidos

Os insights fornecidos pelo gráfico de frequência das vendas por categorias nos oferece uma visão geral das categorias mais populares e destacam a necessidade de ajustes **estratégicos** para atender às demandas do mercado.

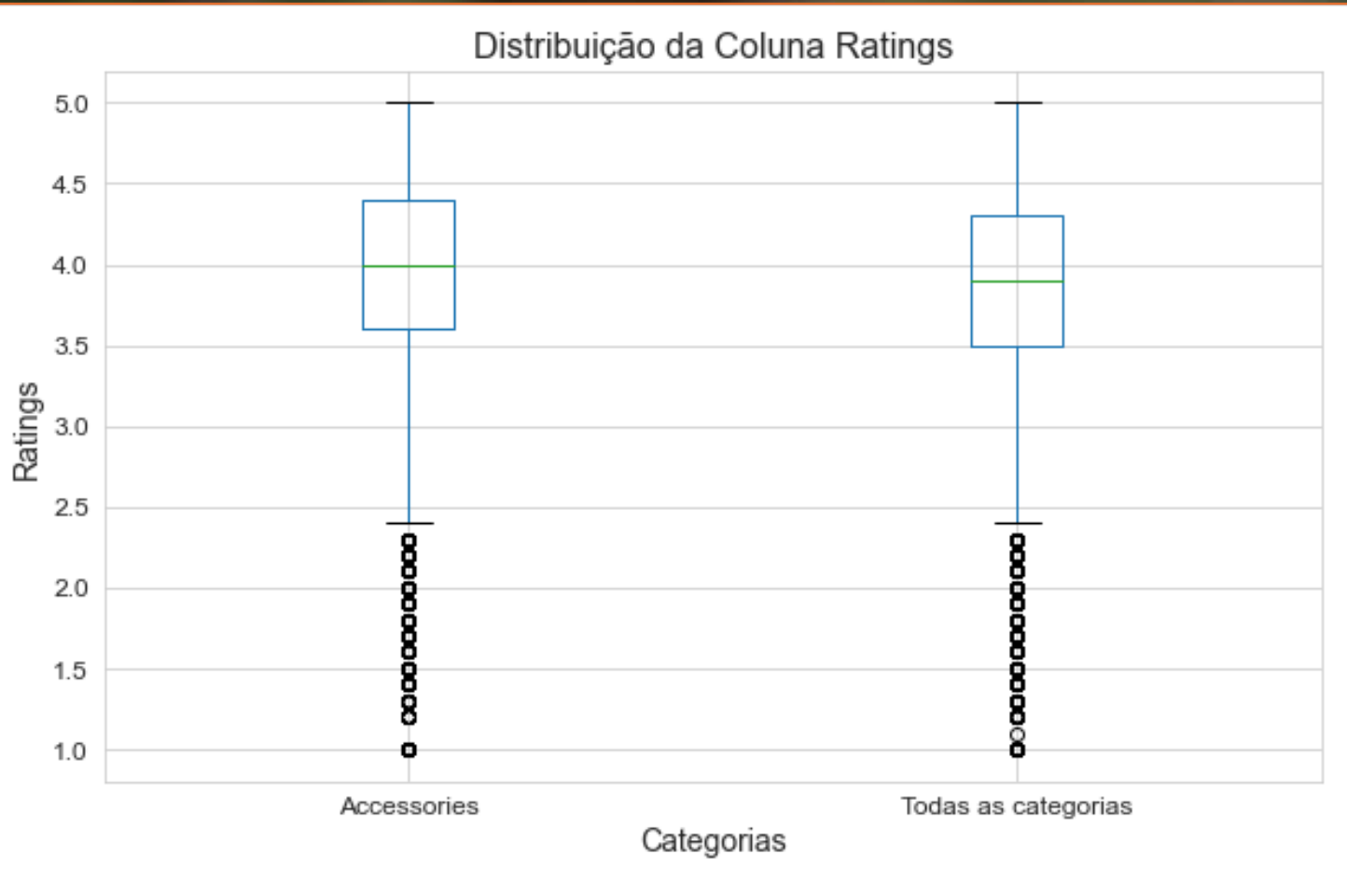
Principais **insights**:

- A categoria "accessories" liderou as vendas com mais de 300.000 unidades, indicando sua popularidade entre os consumidores.
- Tanto "men's clothing" quanto "women's clothing" alcançaram 150.000 vendas, demonstrando que roupas são itens comuns nas compras online.
- "TV, audio e cameras" registrou mais de 100.000 vendas, possivelmente devido ao crescimento da produção de conteúdo online.
- Não houve vendas nas categorias "home, kitchen, pets", "pet supplies", "music" e "grocery e gourmet foods", indicando falta de interesse ou oferta inadequada.



# Visualização dos resultados obtidos

O gráfico de distribuição das avaliações é importante pois ela fornece uma **visão ampla** sobre a qualidade e a consistência dos produtos em cada categoria.

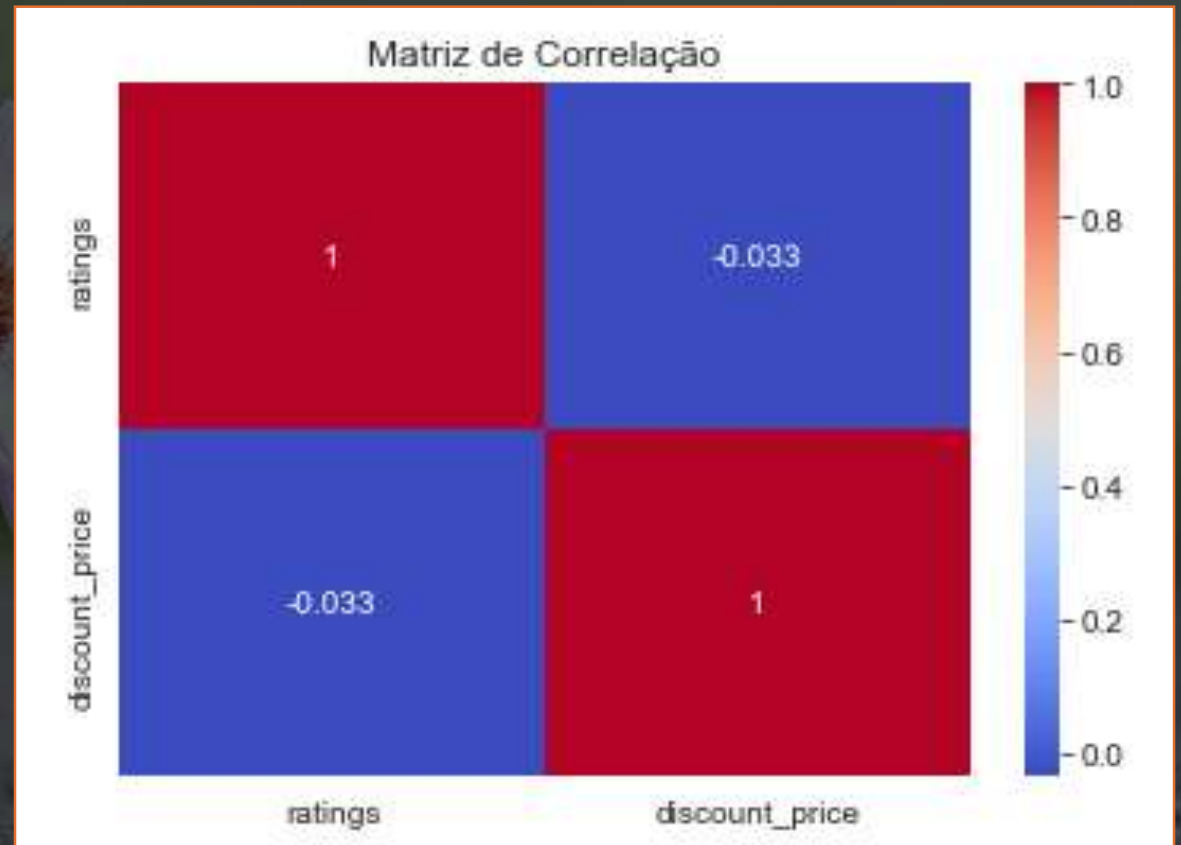


A distribuição das avaliações revela que a categoria "accessories" tem uma maior concentração de ratings próximos a 4.0, indicando **uma qualidade consistente e avaliações semelhantes dos consumidores**. Por outro lado, as outras categorias apresentam uma distribuição mais variada, **sugerindo uma maior diversidade em termos de qualidade e avaliações dos produtos**. Apesar disso, é importante analisar outros aspectos dos dados para confirmar essas observações. Quanto aos outliers, na categoria "accessories" há poucas avaliações abaixo de 2.5, indicando um desempenho geralmente satisfatório.

# Visualização dos resultados obtidos

A análise da matriz de correlação e do mapa de calor revelou uma correlação negativa fraca de -0.033 entre a quantidade de avaliações e o preço do produto com desconto. Isso indica que, em média, **os produtos mais caros não tendem a ter menos avaliações do que os produtos mais baratos**. No entanto, é importante lembrar que a correlação não implica causalidade, e existem outros fatores, como qualidade do produto, marca e popularidade do vendedor, que podem influenciar tanto o preço quanto a quantidade de avaliações. Portanto, é necessário considerar mais informações para entender melhor essa relação complexa

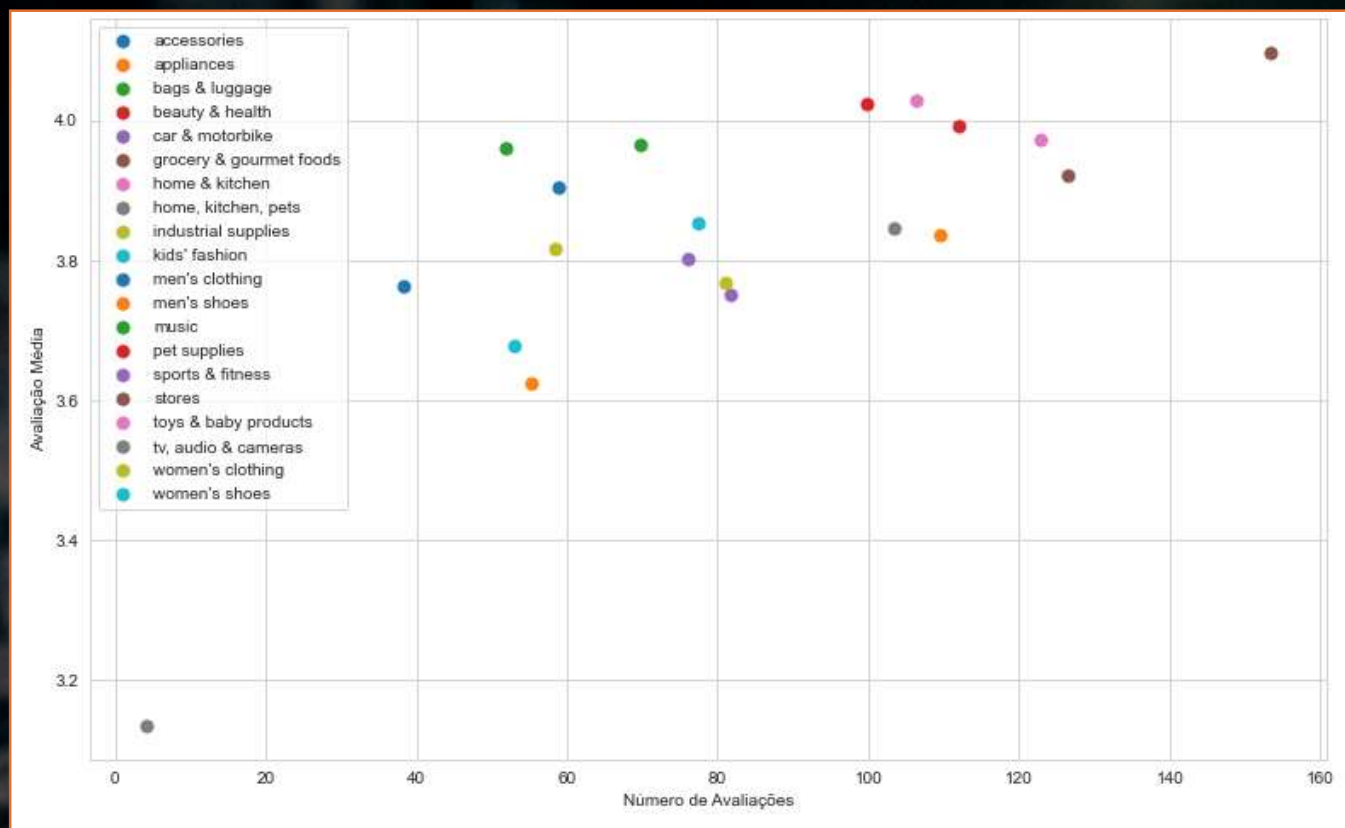
A correlação nos permite entender a **relação** entre a quantidade de avaliações e o preço do produto com desconto.





# Visualização dos resultados obtidos

O gráfico de dispersão mostra a relação entre a **avaliação média** e o **número de avaliações** por categoria de produtos.



Categorias com avaliação média mais alta estão posicionadas acima, indicando maior satisfação. Categorias com mais avaliações estão localizadas à direita, indicando maior participação dos consumidores. O gráfico nos ajuda a identificar relações entre número de avaliações e avaliação média, além de comparar o desempenho das categorias.

- A categoria “stores” é um outlier na avaliação média por quantidade de avaliações.
- A categoria “bags e luggage ” e “music” têm poucas avaliações porém um bom desempenho médio.
- As compras da categoria “grocery e gourmet foods” possuem muitas avaliações mas um desempenho que não acompanha essa quantidade.