Universidade Presbiteriana Mackenzie

PROJETO APLICADO I

ETAPA 2 – APLICANDO CONHECIMENTO

Antônio Alcivan da Silva

Ítalo Pedro dos Santos Pereira

Joyce Rhaellem Alves Costa

Priscila Gabrielly Mendes

Rafael Kazuo Kondo

Universidade Presbiteriana Mackenzie

PROJETO APLICADO I

ETAPA II – APLICANDO CONHECIMENTO

Curso: Faculdade de Computação e Informática – Tecnologia em Ciência de Dados

Semestre: 1º Semestre 2023

Componente Curricular/Tema: Projeto Aplicado I – Aula 02 Aplicando Conhecimento

Grupo 1 – Nome dos Integrantes:

Antônio Alcivan da Silva **TIA:** 22501355
Ítalo Pedro dos Santos Pereira **TIA:** 22502440
Joyce Rhaellem Alves Costa **TIA:** 22514287
Priscila Gabrielly Mendes **TIA:** 22504370
Rafael Kazuo Kondo **TIA:** 22009183

Nome do Professor: LEONARDO MASSAYUKI TAKUNO

Universidade Presbiteriana Mackenzie

Sumário

Glossário

Algoritmos: é uma sequência de instruções ou comandos realizados de maneira sistemática com o objetivo de resolver um problema ou executar uma tarefa.

Aprendizado de máquina: é um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.

Mineração de dados: é uma técnica assistida por computador usada em análises para processar e explorar grandes conjuntos de dados.

Clusters: significa integrar dois ou mais computadores para que trabalhem simultaneamente no processamento de uma determinada tarefa.

Data Science: é um estudo muito disciplinado com relação aos dados e demais informações inerentes à empresa e as visões que cercam um determinado assunto.

Datasets: é um arquivo que pode conter centenas ou até milhares de dados sobre um determinado assunto.

Arquivo CSV (valores separados por vírgulas): é um arquivo de texto com formato específico para possibilitar o salvamento dos dados em um formato estruturado de tabela.

Marketplace: nada mais é do que um portal de e-commerce colaborativo, isto é, um site que reúne ofertas de produtos e serviços de diversos vendedores, como se fosse um shopping virtual.

GitHub: é uma plataforma totalmente online onde você pode criar repositórios e hospedar neles seus projetos, colaborar com softwares open source, seguir programadores e interagir com códigos de terceiros.

Outputs: mecanismo através do qual a informação armazenada e processada num computador é transferida para um meio externo; saída.

Warning: trata-se de mensagens de aviso que são normalmente emitidas em situações em que é útil alertar o usuário sobre alguma condição em um programa.

Outliers: é um dado que se distancia radicalmente dos demais que compõem a amostra analisada.

Objetivo de estudo e problema de pesquisa

O objetivo deste estudo é promover o crescimento do comércio eletrônico por meio da aplicação eficiente de técnicas de clusterização e recomendação de produtos. Para isso, utilizamos uma base de dados de vendas da empresa para avaliarmos a possibilidade de aprimorar o desempenho de vendas por meio dessas técnicas.

A análise dos dados de vendas permite identificar padrões de comportamento dos clientes e suas preferências. Com base nesses dados, é possível aplicar a técnica de clusterização para agrupar produtos em categorias semelhantes, o que proporciona uma melhor compreensão do perfil dos clientes e suas necessidades.

A aplicação dessas técnicas pode aumentar a eficiência do comércio eletrônico, melhorando a experiência do usuário e, consequentemente, impulsionando as vendas. Ademais, a análise dos dados de vendas pode fornecer valiosos insights sobre produtos com maior potencial de crescimento, permitindo que a empresa ajuste sua estratégia de vendas para atender às demandas dos clientes de forma mais eficaz.

Sistemas de Recomendação de produtos: os sistemas de recomendação são algoritmos que coletam informações sobre as preferências de um usuário em um site de compras e usam essas informações para fazer sugestões personalizadas de produtos que ele pode gostar. Eles são amplamente usados em sites de e-commerce para melhorar a experiência do usuário, aumentar a satisfação do cliente e impulsionar as vendas. Os sistemas de recomendação usam técnicas de aprendizado de máquina e mineração de dados para analisar o comportamento do usuário, como histórico de navegação, compras anteriores, avaliações e classificações de produtos, para sugerir itens que possam ser relevantes ou de interesse para o usuário.

Clusterização dos Produtos: a clusterização de produtos para e-commerce é uma técnica de análise de dados que agrupa itens de um catálogo de produtos em clusters, com base em características ou atributos semelhantes. É uma abordagem usada por empresas de e-commerce para melhorar a organização, a categorização e a busca de produtos em seus sites. A clusterização pode ajudar a fornecer uma visão mais clara do comportamento do cliente e identificar padrões de compra. Com essa técnica, é possível identificar grupos de produtos que compartilham características e criar campanhas de marketing direcionadas, otimizar as sugestões de produtos e melhorar a experiência do usuário, levando a um aumento nas vendas e na fidelidade do cliente. A clusterização de produtos é geralmente implementada usando algoritmos de aprendizado de máquina e técnicas de mineração de dados.

Apresentação da Empresa

Nome da empresa

Amazon.com, Inc.

Missão e visão

De acordo com o próprio site de carreira da Amazon, a sua missão é ser “a empresa mais centrada no cliente da Terra”, onde os clientes possam encontrar e descobrir qualquer coisa que queiram comprar online".

Valores:

- **Obsessão pelo cliente:** A empresa se esforça para atender às necessidades dos clientes de maneira excepcional e superar suas expectativas;
- **Propensão para ação:** A Amazon valoriza a tomada de decisão rápida e a implementação ágil de soluções;
- **Inovação:** A empresa busca constantemente inovar e criar soluções que possam melhorar a vida dos clientes;
- **Liderança:** A Amazon busca liderar em todas as áreas em que atua, seja em tecnologia, logística ou em outros aspectos do negócio;
- **Excelência operacional:** A empresa busca operar de maneira eficiente e eficaz, eliminando desperdícios e reduzindo custos para oferecer preços mais baixos aos clientes;
- **Trabalho em equipe:** A Amazon valoriza o trabalho em equipe e a colaboração entre os funcionários para alcançar objetivos comuns.

Segmento de atuação:

O segmento de atuação do comércio online da Amazon é abrangente e inclui diversas categorias de produtos e serviços. Segundo o relatório anual da Amazon de 2020, os principais segmentos de negócios da empresa são:

- **Varejo online:** inclui a venda de produtos físicos para consumidores finais, como eletrônicos, vestuário, alimentos, produtos de beleza, entre outros;
- **Serviços de assinatura:** inclui serviços digitais, como Amazon Prime, Prime Video, Prime Music, entre outros;
- **Serviços em nuvem:** oferece soluções de armazenamento, análise de dados e computação em nuvem para empresas, por meio da Amazon Web Services (AWS);
- **Publicidade:** inclui a venda de espaço publicitário na plataforma da Amazon para anunciantes.

Market Share/posicionamento no mercado

A Amazon é uma das maiores empresas de comércio eletrônico do mundo e tem uma presença significativa em vários mercados. De acordo com dados da eMarketer, a Amazon detém uma participação de mercado de 39,7% no varejo online dos Estados Unidos em 2021, o que representa um aumento em relação aos 38,7% em 2020. A empresa também é líder em serviços de nuvem, com uma participação de mercado de 32% em 2020, de acordo com a Synergy Research Group.

No mercado de publicidade digital, a Amazon também tem aumentado sua participação. Em 2021, a empresa foi a terceira maior plataforma de publicidade digital nos Estados Unidos, com uma participação de mercado de 10,3%, atrás apenas do Google e do Facebook.

Número de colaboradores

De acordo com o relatório anual da Amazon.com, Inc. de 2020, a empresa empregava mais de 1,3 milhão de funcionários em todo o mundo. Esses funcionários trabalham em diversas áreas, incluindo varejo online, tecnologia, logística, serviços em nuvem, entre outros. É importante notar que esse número de funcionários pode ter mudado desde a publicação do relatório, pois a Amazon continua crescendo e expandindo suas operações.

Iniciativas na área de Data Science

A Amazon tem investido significativamente em iniciativas de Data Science, com o objetivo de utilizar dados para melhorar seus produtos, serviços e operações. Algumas das iniciativas da empresa na área de Data Science incluem:

1. **Amazon Machine Learning:** A Amazon oferece uma plataforma de aprendizado de máquina baseada em nuvem que permite que os desenvolvedores construam e implantem modelos de aprendizado de máquina em escala.
   A plataforma é integrada aos serviços da AWS e fornece recursos de automatização de processos para simplificar o processo de treinamento de modelos;

2. **Amazon Personalize:** A Amazon Personalize é uma plataforma de aprendizado de máquina que permite que as empresas criem recomendações personalizadas para seus clientes. A plataforma utiliza algoritmos de aprendizado de máquina para analisar dados de clientes, incluindo histórico de compras e comportamento de navegação, para criar recomendações personalizadas em tempo real;

3. **Amazon Forecast:** A Amazon Forecast é uma plataforma de previsão baseada em nuvem que permite que as empresas criem previsões precisas para seus negócios. A plataforma utiliza algoritmos de aprendizado de máquina para analisar dados históricos e criar modelos de previsão que podem ser usados para prever vendas futuras, demanda de produtos e outros fatores críticos de negócios.

Trabalhos em destaque

A Amazon é uma empresa que atua em diversos segmentos e tem inúmeros projetos em andamento. Aqui estão alguns trabalhos em destaque da Amazon em diferentes áreas, juntamente com suas referências bibliográficas:

1. **Alexa Conversations:** A equipe de Alexa Conversations da Amazon está trabalhando em um sistema de diálogo conversacional que permite que os usuários interajam com a assistente virtual da Amazon, Alexa, de maneira mais natural e intuitiva. O sistema utiliza técnicas de aprendizado de máquina e permite que os usuários realizem tarefas mais complexas por meio de uma única interação com a assistente virtual;

2. **Amazon Go:** A Amazon Go é uma rede de lojas sem caixas registradoras que utilizam tecnologia de visão computacional e aprendizado de máquina para permitir que os clientes façam compras e saiam da loja sem precisar passar pelo processo de checkout. A empresa abriu sua primeira loja em 2018 e desde então tem expandido sua rede de lojas em vários países;

3. **Amazon Web Services:** A Amazon Web Services (AWS) é a divisão de serviços em nuvem da Amazon, que fornece uma ampla variedade de serviços de computação em nuvem, incluindo armazenamento, processamento, análise de dados, machine learning e muito mais.
   A AWS é líder em serviços de nuvem e tem sido responsável por grande parte dos lucros da Amazon nos últimos anos.

Problema do estudo

O que falta?

Quais foram os critérios e os algoritmos utilizados para a clusterização e a recomendação de produtos?

Quais foram as métricas e os indicadores utilizados para avaliar o desempenho das técnicas?

O que incomoda?

Concorrência e regulação: a empresa enfrenta uma forte concorrência de outras empresas de tecnologia e de comércio eletrônico, como Google, Facebook, Apple, Microsoft, Alibaba e Walmart. Além disso, a empresa também está sob o escrutínio de órgãos reguladores e governos de vários países, que investigam possíveis práticas anticompetitivas, violações de privacidade e abusos de poder da empresa.

Reclamações dos consumidores: a empresa tem recebido muitas reclamações dos consumidores sobre atrasos na entrega, problemas com os produtos, dificuldades no atendimento e na devolução, entre outros. Segundo o site Reclame Aqui, a empresa tem uma reputação de 8.4 em 10, mas recebeu mais de 44 mil reclamações nos últimos 6 meses, sendo que cerca de 3 mil não foram respondidas.

O que pode ser melhorado?

Comparar as técnicas de clusterização e recomendação de produtos utilizadas com outras alternativas disponíveis na literatura ou no mercado, para avaliar a sua eficácia e eficiência;

Realizar testes com diferentes cenários e variáveis, para verificar a robustez e a sensibilidade das técnicas aplicadas;

Apresentar exemplos concretos e ilustrativos de como as técnicas de clusterização e recomendação de produtos podem melhorar a experiência do usuário e o crescimento de vendas da empresa Amazon.

Qual é o gap?

- **Estado atual:** a empresa Amazon utiliza técnicas de clusterização e recomendação de produtos para melhorar o seu comércio eletrônico, mas não se sabe quão efetivas e eficientes são essas técnicas e como elas podem ser otimizadas;
- **Estado desejado:** a empresa Amazon quer melhorar o seu crescimento de vendas por meio de técnicas de clusterização e recomendação de produtos que sejam capazes de identificar o perfil dos clientes e suas preferências, agrupar os produtos em categorias semelhantes e sugerir itens relevantes ou de interesse para o usuário;
- **GAP:** como utilizar técnicas de clusterização e recomendação de produtos para melhorar o crescimento de vendas da empresa Amazon?

Há um padrão que pode ser observado?

Nessa análise inicial dos dados, conseguimos observar que a grande maioria dos produtos consistem em eletrônicos, computadores & acessórios e eletrodomésticos, com avaliações dadas pelos consumidores variando entre 3,5 e 4,5 (com 5,0 sendo a avaliação máxima).

Há uma afirmação que pode ser contestada?

Até o momento da análise ainda não podemos contestar nenhuma afirmação.

Metadados

Objetivo

Caracterizar e registrar os datasets que temos para o estudo.

Tipo de arquivo (csv, json, xml etc.)

O arquivo dataset é um CSV

Origem dos dados (aberto ou privado)

Os dados são de origem pública, presentes em produtos no próprio marketplace da empresa

Sensibilidade (possui dados sensíveis)

Não há sensibilidade nos dados, já que se trata somente das avaliações públicas dos produtos

Validade (quando foi gerado, quando se torna obsoleto)

Os dados não expiram

Proprietário do dado (quem é responsável pelo dataset)

KARKAVELRAJA J foi responsável pelo processo de ETL (extrair, transformar e carregar) da plataforma da Amazon e por disponibilizar na plataforma Kaggle

Descrição dos atributos (definição e tipo de dado)

product_id - ID do produto (STR)

product_name - Nome do produto (STR)

category - Categoria do produto (STR)

discounted_price - Preço com desconto do produto (FLOAT)

actual_price - Preço real do produto (FLOAT)

discount_percentage - Percentual de desconto do produto (FLOAT)

rating - Avaliação do produto (FLOAT)

rating_count - Número de pessoas que votaram na avaliação da Amazon (FLOAT)

about_product - Descrição do produto (STR)

user_id - ID do usuário que escreveu a avaliação do produto (STR)

user_name - Nome do usuário que escreveu a avaliação do produto (STR)

review_id - ID da avaliação do usuário (STR)

review_title - Avaliação curta do produto (STR)

review_content - Avaliação longa do produto (STR)

img_link - Link da imagem do produto (STR)

product_link - Link do site oficial do produto (STR)

Análise Exploratória de Dados

Os dados de origem e análise feita se encontram dentro do link do GitHub a seguir: https://github.com/italopspereira/PROJE_APLIC_I/tree/main/2%20-%20Etapa

Segue os principais outputs referentes a análise de dados. Vale lembrar, que alguns outputs são extensos e não couberam dentro da tela. Existem algumas mensagens de warning relacionadas a parâmetros deprecados, porém, estes não devem influenciar no resultado final.

Número de exemplares (linhas) e dimensões (colunas)

In [ ]:
```
# Todas as colunas presentes no dataset
df.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_title',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
# Total de linhas e colunas
df.shape
```

Out[ ]:
```
(1465, 16)
```

Tipos de dados

In [ ]:
```
df.dtypes
```

Out[ ]:
```
product_id              object
product_name            object
category                object
discounted_price        object
actual_price            object
discount_percentage     object
rating                  object
rating_count            object
about_product           object
user_id                 object
user_name               object
review_id               object
review_title            object
review_content          object
img_link                object
product_link            object
dtype: object
```

Medidas de posição e dispersão

In [ ]:
```
# Média
df.mean()
```

```
<ipython-input-29-c61f0c8f89b5>:1: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.mean()
```

Out[ ]:
```
discounted_price       3125.310874
actual_price           5444.990635
discount_percentage       0.476915
rating                    4.096519
rating_count          18295.541353
dtype: float64
```

In [ ]:
```
# Mediana
df.median()
```

```
<ipython-input-30-c6e0c62a3834>:2: FutureWarning: The default value of numeric_only in DataFrame.median is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.median()
```

Out[ ]:
```
discounted_price       799.0
actual_price          1650.0
discount_percentage      0.5
rating                   4.1
rating_count          5179.0
dtype: float64
```

In [ ]:
```
# Quartis (primeiro quartil)
df.quantile(q=0.25)
```

```
<ipython-input-31-3f00c7a07290>:2: FutureWarning: The default value of numeric_only in DataFrame.quantile is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.quantile(q=0.25)
```

Out[ ]:
```
discounted_price       325.00
actual_price           800.00
discount_percentage      0.32
rating                   4.00
rating_count          1186.00
Name: 0.25, dtype: float64
```

In [ ]:
```
# Quartis (terceiro quartil)
df.quantile(q=0.75)
```

```
<ipython-input-33-048d2407fab8>:2: FutureWarning: The default value of numeric_only in DataFrame.quantile is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.quantile(q=0.75)
```

Out[ ]:
```
discounted_price       1999.00
actual_price           4295.00
discount_percentage       0.63
rating                    4.30
rating_count          17336.50
Name: 0.75, dtype: float64
```

In [ ]:
```
# Moda
print(df['discounted_price'].mode())
print(df['actual_price'].mode())
print(df['discount_percentage'].mode())
print(df['rating'].mode())
print(df['rating_count'].mode())
```

```
0    199.0
Name: discounted_price, dtype: float64
0    999.0
Name: actual_price, dtype: float64
0    0.5
1    0.6
Name: discount_percentage, dtype: float64
0    4.1
Name: rating, dtype: float64
0    9378.0
Name: rating_count, dtype: float64
```

In [ ]:
```
# Amplitude
print('discounted_price :',df['discounted_price'].max() - df['discounted_price'].min())
print('actual_price :',df['actual_price'].max() - df['actual_price'].min())
print('discount_percentage :',df['discount_percentage'].max() - df['discount_percentage'].min())
print('rating :',df['rating'].max() - df['rating'].min())
print('rating_count :',df['rating_count'].max() - df['rating_count'].min())
```

```
discounted_price : 77951.0
actual_price : 139861.0
discount_percentage : 0.94
rating : 3.0
rating_count : 426971.0
```

In [ ]:
```
# Variância
print('discounted_price:', round(df['discounted_price'].var(), 3))
print('actual_price :', round(df['actual_price'].var(), 3))
print('discount_percentage :', round(df['discount_percentage'].var(), 3))
print('rating :', round(df['rating'].var(), 3))
print('rating_count :', round(df['rating_count'].var(), 3))
```

```
discounted_price: 48223363.522
actual_price : 118261859.323
discount_percentage : 0.047
rating : 0.085
rating_count : 1827892968.355
```

In [ ]:
```
# Desvio Padrão
print('discounted_price:', round(df['discounted_price'].std(), 3))
print('actual_price :', round(df['actual_price'].std(), 3))
print('discount_percentage :', round(df['discount_percentage'].std(), 3))
print('rating :', round(df['rating'].std(), 3))
print('rating_count :', round(df['rating_count'].std(), 3))
```

```
discounted_price: 6944.304
actual_price : 10874.827
discount_percentage : 0.216
rating : 0.292
rating_count : 42753.865
```

In [ ]:
```
# Covariância
print(df.cov())
```

```
                   discounted_price  actual_price  discount_percentage  \
discounted_price        4.822336e+07  7.264202e+07          -364.215122
actual_price            7.264202e+07  1.182619e+08          -277.867820
discount_percentage    -3.642151e+02 -2.778678e+02             0.046811
rating                  2.437229e+02  3.858887e+02            -0.009774
rating_count           -8.098244e+06 -1.681132e+07           108.076044

                         rating  rating_count
discounted_price      243.722892 -8.098244e+06
actual_price          385.888738 -1.681132e+07
discount_percentage    -0.009774  1.080760e+02
rating                  0.085022  1.266013e+03
rating_count         1266.013169  1.827893e+09
<ipython-input-54-0a12e4c3650a>:1: FutureWarning: The default value of numeric_only in DataFrame.cov is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  print(df.cov())
```

Distribuição e frequência

In [ ]:
```
frequencia = df['actual_price'].value_counts()
porcentagem = df['actual_price'].value_counts(normalize = True) * 100
```

In [ ]:
```
dist_freq = pd.DataFrame({'Frequência':frequencia, 'Porcentagem %':porcentagem})

with pd.option_context('display.max_rows', None,
                       'display.max_columns', None,
                       'display.precision', 3,
                       ):
  print(dist_freq)
```

```
          Frequência  Porcentagem %
999.00           120          8.191
499.00            71          4.846
1999.00           56          3.823
1499.00           37          2.526
399.00            34          2.321
599.00            33          2.253
699.00            29          1.980
799.00            24          1.638
2999.00           22          1.502
1299.00           21          1.433
899.00            20          1.365
9999.00           19          1.297
2499.00           19          1.297
1099.00           19          1.297
4999.00           17          1.160
5999.00           17          1.160
3999.00           17          1.160
299.00            17          1.160
1599.00           17          1.160
```

Correlações

In [ ]:
```
df.corr()
```

```
<ipython-input-27-2f6f6606aa2c>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.corr()
```

Out[ ]:

	discounted_price	actual_price	discount_percentage	rating	rating_count
discounted_price	1.000000	0.961915	-0.242412	0.120365	-0.027261
actual_price	0.961915	1.000000	-0.118098	0.121695	-0.036137
discount_percentage	-0.242412	-0.118098	1.000000	-0.154924	0.011691
rating	0.120365	0.121695	-0.154924	1.000000	0.102318
rating_count	-0.027261	-0.036137	0.011691	0.102318	1.000000

Valores perdidos ou incorretos

Os valores descartados foram 1 - |, 2 - Linhas vazias(literalmente vazias), e todas os caracteres da moeda indiana que estava presente em toda a coluna de preço atual

Anomalias e outliers

Utlizado o Método IQR (Interquartile Range): O método IQR é uma técnica simples e amplamente utilizada para identificar outliers em um conjunto de dados. O primeiro passo é calcular o intervalo interquartil (IQR), que é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1). Em seguida, os outliers são definidos como quaisquer pontos abaixo de Q1 - 1,5 * IQR ou acima de Q3 + 1,5 * IQR.

In [ ]:

```
# Calcula Q1, Q3 e IQR
Q1 = df['actual_price'].quantile(0.25)
Q3 = df['actual_price'].quantile(0.75)
IQR = Q3 - Q1

outliers = pd.DataFrame(df[(df['actual_price'] < Q1 - 1.5*IQR) | (df['actual_price'] > Q3 + 1.5*IQR)])

with pd.option_context('display.max_rows', None,
                       'display.max_columns', None,
                       'display.precision', 3,
                       ):
  print(outliers['actual_price'])
```

```
16       24999.0
19       21990.0
22       22900.0
24       19990.0
26       19999.0
38       45999.0
41       34999.0
53       12999.0
57       21999.0
61       47900.0
64       24999.0
67       14990.0
72       42999.0
77       30990.0
```

Referências Bibliográficas

https://aws.amazon.com/machine-learning/
https://www.amazon.com/b?ie=UTF8&node=16008589011
https://aws.amazon.com/about-aws/
https://www.aboutamazon.com/about-us
https://www.sec.gov/Archives/edgar/data/1018724/000101872421000014/amzn-20201231xex101.htm
https://influencermarketinghub.com/amazon-statistics/#toc-1
https://techcrunch.com/2022/04/29/amazon-still-undisputed-king-of-public-cloud-but-microsoft-is-creeping-closer/
https://www.insiderintelligence.com/content/amazon-dominates-us-ecommerce-though-its-market-share-varies-by-category
https://www.amazon.science/blog/how-alexa-is-learning-to-converse-more-naturally