

ANÁLISE COMPORTAMENTAL DE CONSUMO DOS USUÁRIOS

Laura Damaceno de Almeida





OBJETIVO

Entender os diversos comportamentos de consumo dos usuários



HIPÓTESES A SEREM VALIDADAS



H0:

Não há grupos distintos de usuários, pois há grupos com interesses mistos e faz sentido ter um mix de temas

H1:

Há grupos distintos de usuários e por conta disso deveriam receber conteúdos distintos



METODOLOGIA UTILIZADA

01

ENTENDIMENTO DOS DADOS

Etapa para entendimento do formato do dados e análise qualitativa

Entendimento do perfil dos usuários e análise de consumo

ANÁLISE DESCRITIVA E DIAGNÓSTICA

02

03

PRÉ-PROCESSAMENTO

Transformação aplicada nas variáveis numéricas para garantir que todas sigam o mesmo intervalo de valores

Execução do algoritmo de segmentação para avaliar as hipóteses

SEGMENTAÇÃO E ANÁLISE DOS RESULTADOS

04



A person with long, dark, curly hair is seen from the side, wearing large, over-ear headphones. They are holding a smartphone in front of them, looking at the screen. The phone's screen displays a music player interface with a large circular album art and some text. The background is dark and out of focus, with a person in a light-colored shirt visible in the distance. The overall lighting is dim, with a blue and purple hue.

ANÁLISE QUALITATIVA



VISÃO GERAL DAS VARIÁVEIS DISPONÍVEIS

Quantidade de usuários: 161.757

Colunas: 44

Quantidade de variáveis qualitativa nominais: 3

Quantidade de variáveis quantitativa contínuas: 37

Quantidade de variáveis quantitativa discretas: 4



ANÁLISE QUALITATIVA

Foi identificado que apenas essas variáveis contém valores nulos. Todavia o percentual de representatividade é bem baixo, portanto para manter a fidelidade dos dados foram desconsiderados os usuários que têm essas informações faltantes

index	colunas	tipo	Qtde valores NaN	% valores NaN	valores únicos por feature
avg_tempo_sessao	avg_tempo_sessao	float64	20	0.000124	161711
razao_sessao_g1_home	razao_sessao_g1_home	float64	2578	0.015937	45629
razao_sessao_ge_home	razao_sessao_ge_home	float64	11784	0.072850	57355
razao_sessao_gshow_home	razao_sessao_gshow_home	float64	15580	0.096317	37636
address_state	address_state	object	7041	0.043528	29
idade	idade	float64	3635	0.022472	110
gender	gender	object	5474	0.033841	4



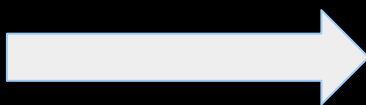
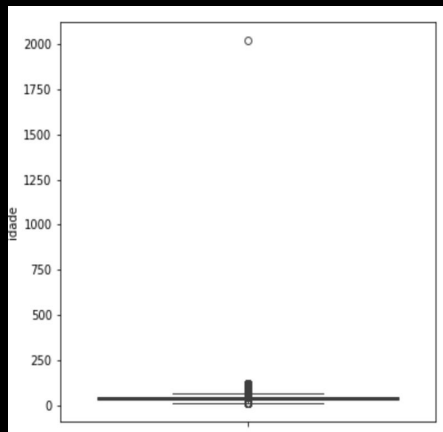
ANÁLISE QUALITATIVA

Além disso, as variáveis: idade, clicks totais, pageviews e sessões estão em um formato errado (deveria ser numérico quantitativo do que contínuo), portanto foram tratadas para não ter problemas durante as análises.

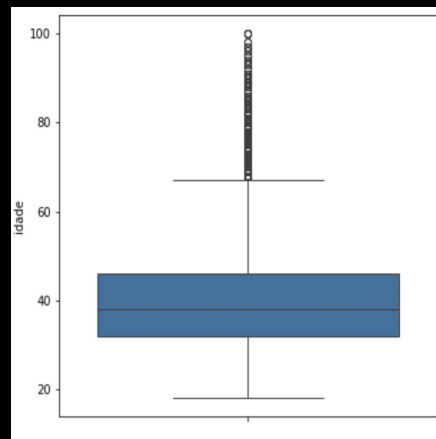


ANÁLISE QUALITATIVA

Outro ponto relevante é que a variável idade apresenta alguns outliers. Há registros de usuários com idades acima de 100 e outros com idade abaixo de 18 anos. Para termos uma análise mais assertiva de clientes elegíveis para os serviços da Globo, foi aplicado um filtro para desconsiderar registros com idade acima de 100 e abaixo de 18 anos.



Após a aplicação
do filtro



METODOLOGIA UTILIZADA

01

ENTENDIMENTO DOS DADOS

Etapa para
entendimento do
formato do dados e
análise qualitativa

Entendimento do perfil
dos usuários e análise de
consumo

ANÁLISE DESCRITIVA E DIAGNÓSTICA

02

03

PRÉ-PROCESSAMENTO

Transformação aplicada
nas variáveis numéricas
para garantir que todas
sigam o mesmo intervalo
de valores

Execução do algoritmo de
segmentação para avaliar
as hipóteses

SEGMENTAÇÃO E ANÁLISE DOS RESULTADOS

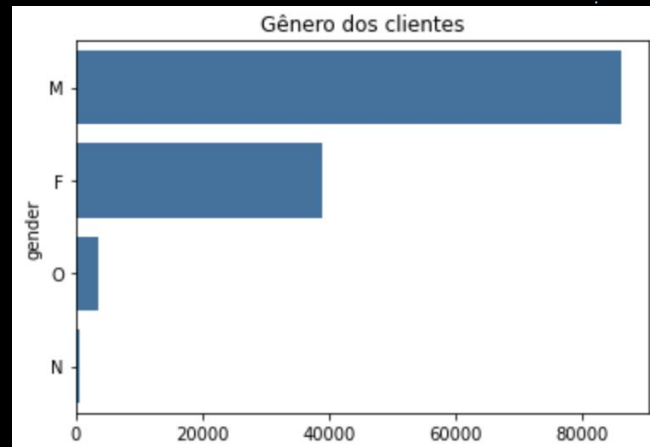
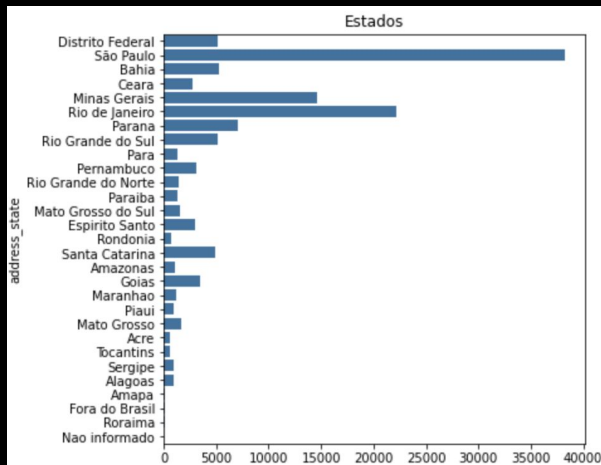
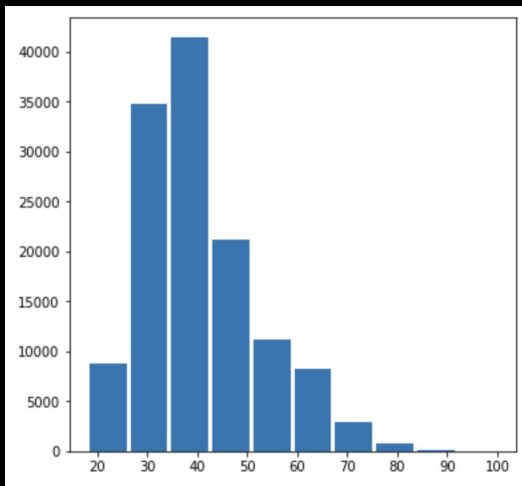
04





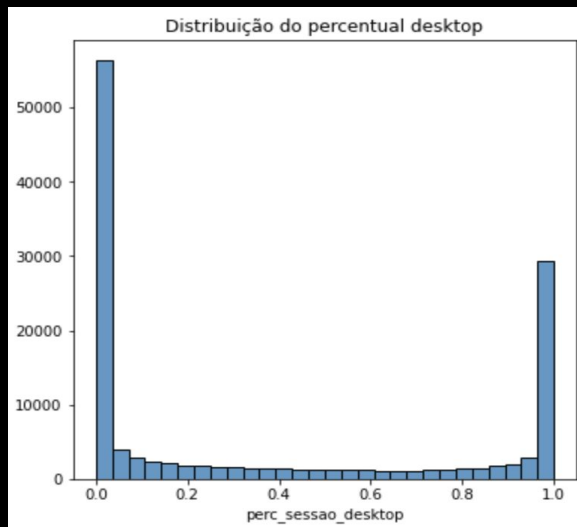
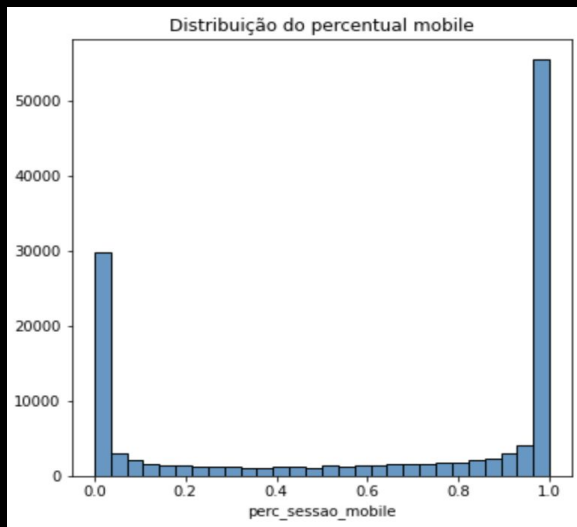
ENTENDIMENTO DO PERFIL E CONSUMO DOS USUÁRIOS

INFORMAÇÕES DEMOGRÁFICAS DOS USUÁRIOS



Usuários têm uma concentração maior entre 30 e 40 anos. Maioria é da região Sudeste do Brasil (São Paulo, Rio de Janeiro e Minas Gerais). Gênero dominante na base de dados são homens seguido de mulheres

USABILIDADE



Pode-se perceber que há um percentual maior de pessoas que sempre usam celular e baixo percentual de pessoas que sempre usam o computador.

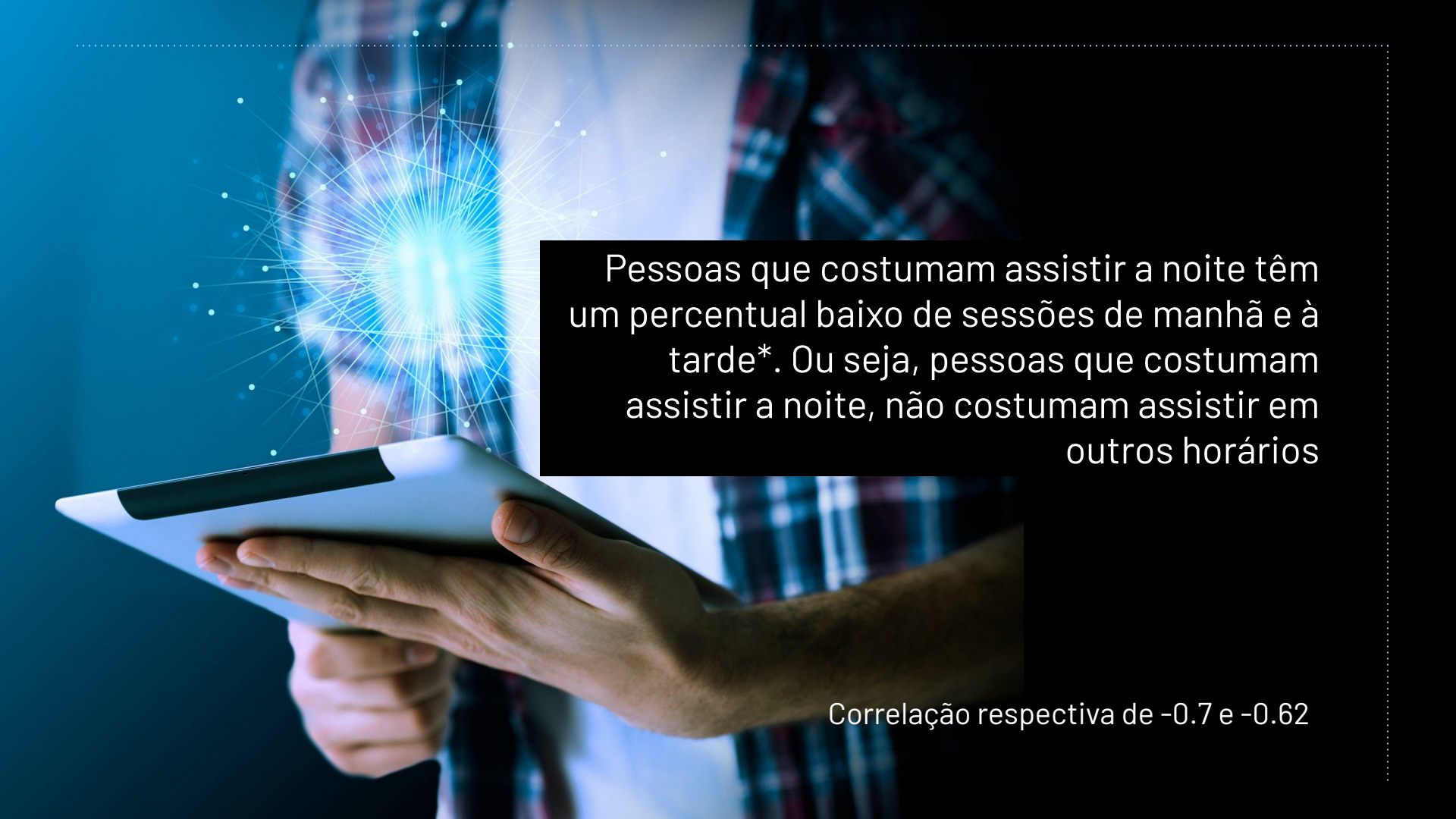
Computador não aparenta ser a primeira opção para as pessoas



PERÍODO DAS SESSÕES

- Em média 33% das sessões dos usuários foram acessado na parte da tarde;
- Em média 30% das sessões dos usuários foram acessado na parte da noite;
- Em média 28% das sessões dos usuários foram acessado na parte da manhã;
- Período da madrugada tem uma média bem baixa (0.07). Ou seja, apenas 7% das sessões ocorrem de madrugada.



A person wearing a plaid shirt is holding a tablet. Overlaid on the image is a glowing blue network graphic with many nodes and connecting lines. A black text box is positioned on the right side of the image.

Pessoas que costumam assistir a noite têm um percentual baixo de sessões de manhã e à tarde*. Ou seja, pessoas que costumam assistir a noite, não costumam assistir em outros horários

Correlação respectiva de -0.7 e -0.62

UAU!

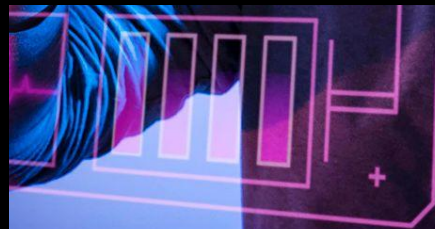
Jornalismo, Esporte, Entretenimento
e Outros foram os temas mais
acessados pelos usuários



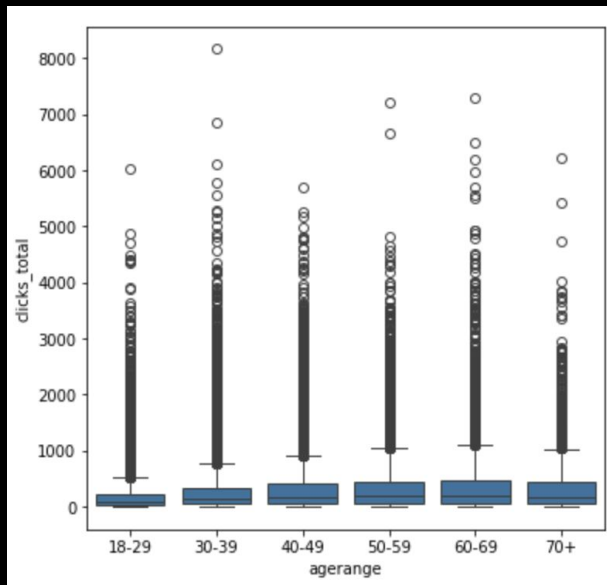
A person is shown from the chest up, wearing a VR headset. The background is a dark, futuristic interface with various HUD elements. At the top, it says 'FUTURISTIC HUD'. Below that, there's a 'TEST' label with a user icon. There are two circular progress indicators, one showing '10%' and the other '90%'. A bar chart is visible on the right. The overall color scheme is dark with blue and purple highlights.

HÁ UMA OPORTUNIDADE DE MELHORIA NOS TEMAS: VALOR, RECEITA, TECHTUDO E GLOBOPLAY

Foi identificado poucos acessos das pessoas durante suas sessões nos temas, tendo um percentual médio de acesso próximo de 0%.



USABILIDADE

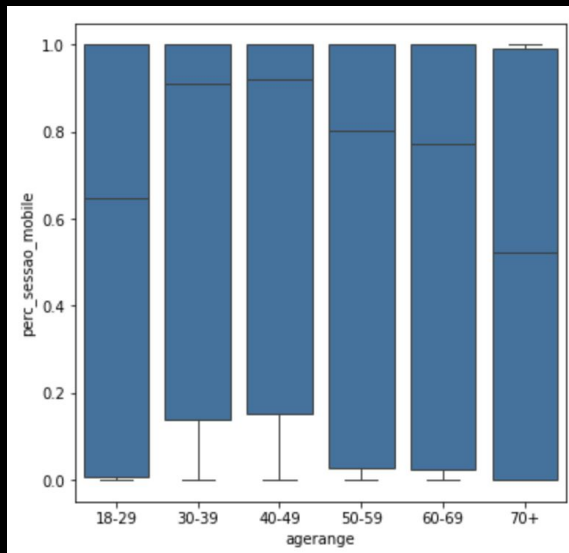


Pode-se perceber que entre 18 e 29 anos há uma distribuição menor de quantidade de cliques necessárias, inclusive a mediana é a menor que todos.

Entretanto 50 e 69 anos tem uma distribuição bem parecida, cuja mediana é maior do que os outros intervalos.



PREFERÊNCIA NO USO DE DISPOSITIVOS POR IDADE

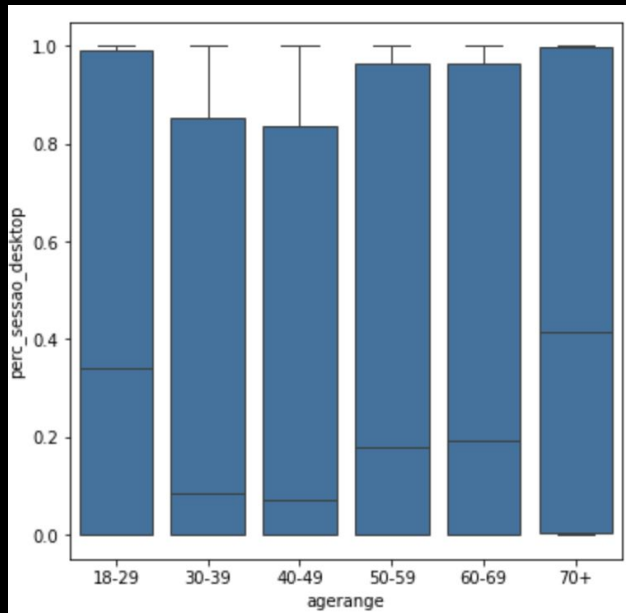


Quem tem mais sessões pelo celular são pessoas entre 30 e 49 anos. A mediana está próxima do Q3 sugere que a distribuição dos dados é assimétrica e tende a ser mais inclinada para o lado direito (ou seja, há mais valores maiores do que menores).

Um outro ponto interessante é que a mediana das pessoas acima de 70 anos, ela é a mais baixa dentre as outras idades.



PREFERÊNCIA NO USO DE DISPOSITIVOS POR IDADE

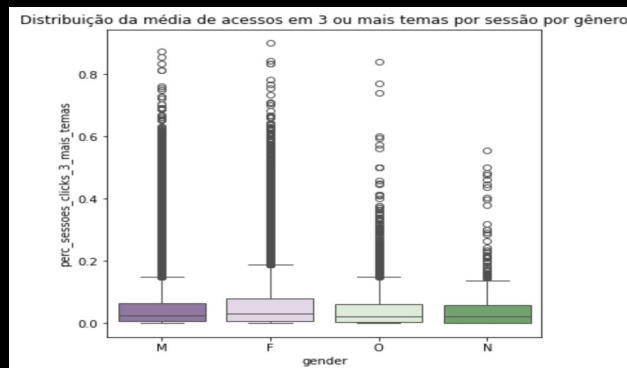
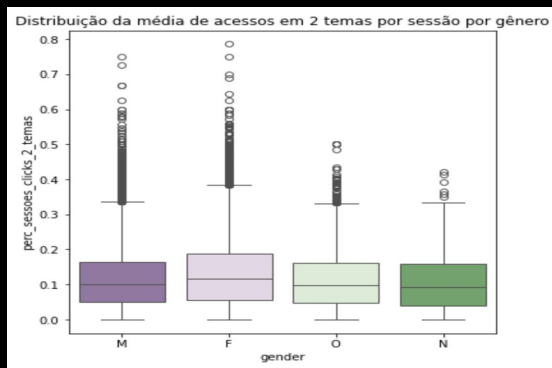


Ao contrário do celular, o computador acaba não sendo a principal referência para pessoas entre 30 e 49 anos.

Todavia apresenta um percentual maior entre pessoas acima de 70 anos e jovens entre 18 e 29 anos.



MULHERES TÊM UMA DISTRIBUIÇÃO MAIOR DE ACESSO A TEMAS DIVERSOS



A partir dos boxplots pode-se visualizar que mulheres têm uma distribuição dos valores acima dos outros gêneros. Foi feito um teste estatístico, através do Teste T, para validar esse insight e foi confirmado que as mulheres têm uma quantidade maior de acessos em diversos temas do que os outros gêneros.





CONSUMO DOS USUÁRIOS

01

ACESSO MADRUGADA

Pessoas que têm sessões de madrugada costumam acessar mais pelo celular (correlação de 0.45)

02

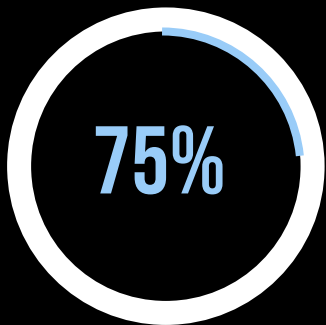
CELULAR

Foi identificado que quanto mais a pessoa acessa pelo celular menos sessões ela tem pelo computador (correlação de -0.99)

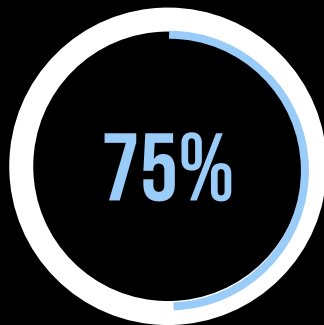
03

ACESSO A TARDE

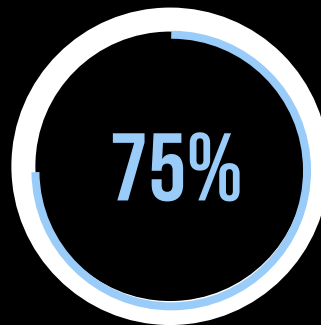
Pessoas que têm sessões à tarde costumam usar o computador (correlação de 0.45)



Dos usuários retornam ao menos **5 vezes na semana** e **mensalmente 17 vezes**.



Dos usuários passam aproximadamente **3 horas na sessão**



Dos usuários têm uma quantidade de cliques totais abaixo de 284





PESSOAS COM INTERESSE EM JORNALISMO E ENTRETENIMENTO TÊM INTERESSE EM DIVERSOS TEMAS*

Através do cálculo de correlação foi identificado que pessoas que tem um percentual de interesse em jornalismo e entretenimento tem correlação forte com clicks em mais de dois temas.



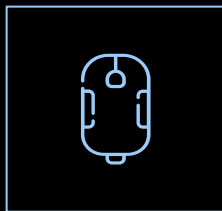
*Entretenimento correlação 0.59 com clicks em mais de 2 temas e 0.56 em mais de 3 temas. Jornalismo correlação com clicks em mais de 2 temas 0.67 e 0.59 com clicks em mais de 3 temas

QUAIS VARIÁVEIS TÊM RELAÇÃO COM O RETORNO SEMANAL?



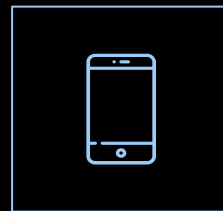
TELAS VISUALIZADAS

Quantidade de telas visualizadas tem uma correlação positiva com o retorno semanal do usuários



CLIQUEs

Quantidade de cliques que a pessoa dá tem uma correlação positiva com o retorno.



SESSÕES

Quanto mais a pessoa retorna semanalmente maior a quantidade de sessões ela tem



A person wearing a VR headset is shown in profile, looking towards the right. Overlaid on the image are various futuristic, glowing blue and green digital interface elements (HUD). These include circular progress indicators with percentages like '10%' and '90%', a bar chart, a line graph, and text labels such as 'FUTURISTIC HUD', 'TEST', and '2-0-2'. The overall aesthetic is high-tech and digital.

PESSOAS COM PERCENTUAL DE SESSÕES NO TEMA OUTROS, TEM INTERESSE EM CONTEÚDOS DE VÍDEO

Foi identificado uma correlação forte (0.72) entre as pessoas com sessões no tema outros e os clique em componentes de vídeo.

METODOLOGIA UTILIZADA

01

ENTENDIMENTO DOS DADOS

Etapa para entendimento do formato do dados e análise qualitativa

Entendimento do perfil dos usuários e análise de consumo

ANÁLISE DESCRITIVA E DIAGNÓSTICA

02

03

PRÉ-PROCESSAMENTO

Transformação aplicada nas variáveis numéricas para garantir que todas sigam o mesmo intervalo de valores

Execução do algoritmo de segmentação para avaliar as hipóteses

SEGMENTAÇÃO E ANÁLISE DOS RESULTADOS

04



METODOLOGIA UTILIZADA

01

ENTENDIMENTO DOS DADOS

Etapa para entendimento do formato do dados e análise qualitativa

Entendimento do perfil dos usuários e análise de consumo

ANÁLISE DESCRITIVA E DIAGNÓSTICA

02

03

PRÉ-PROCESSAMENTO

Transformação aplicada nas variáveis numéricas para garantir que todas sigam o mesmo intervalo de valores

Execução do algoritmo de segmentação para avaliar as hipóteses

SEGMENTAÇÃO E ANÁLISE DOS RESULTADOS

04



HIPÓTESES QUE FORAM INVESTIGADAS

H0

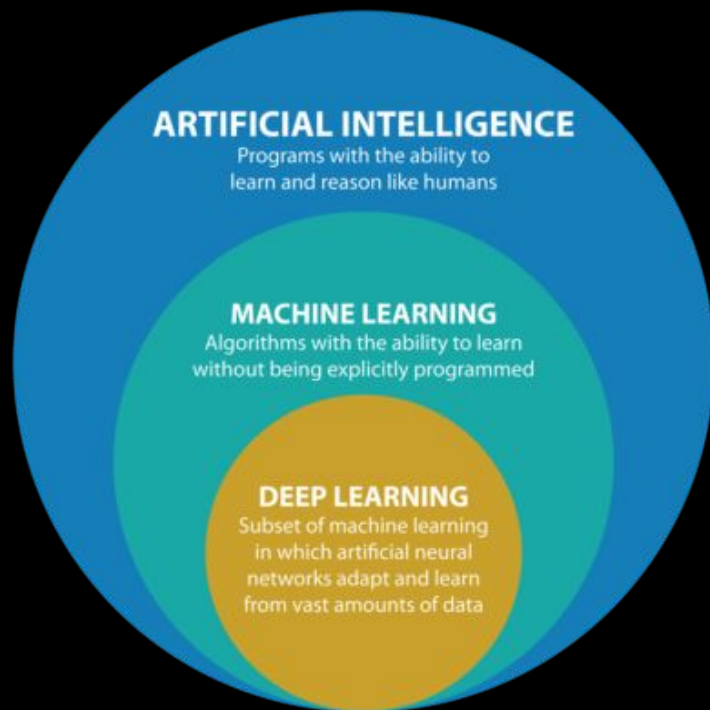
Não há grupos distintos de usuários, pois há grupos com interesses mistos e faz sentido ter um mix de temas

H1

Há grupos distintos de usuários e por conta disso deveriam receber conteúdos distintos



CONCEITOS IMPORTANTES



Para alcançarmos o objetivo de identificar a quantidade de grupos distintos dentro da base, será considerado um algoritmo de Machine Learning (em português, Aprendizado de Máquina).

O K-means é o algoritmo mais utilizado para segmentação de clientes



FUNCIÓNAMENTO K-MEANS

Com base no conjunto de dados e nas características dos usuários, o algoritmo irá agrupá-los em k grupos



FUNCIONAMENTO K-MEANS

Por exemplo: se k for igual a 3. Então o algoritmo vai agrupar os usuários em 3 grupos diferentes com base nas suas características:



FUNCIONAMENTO K-MEANS

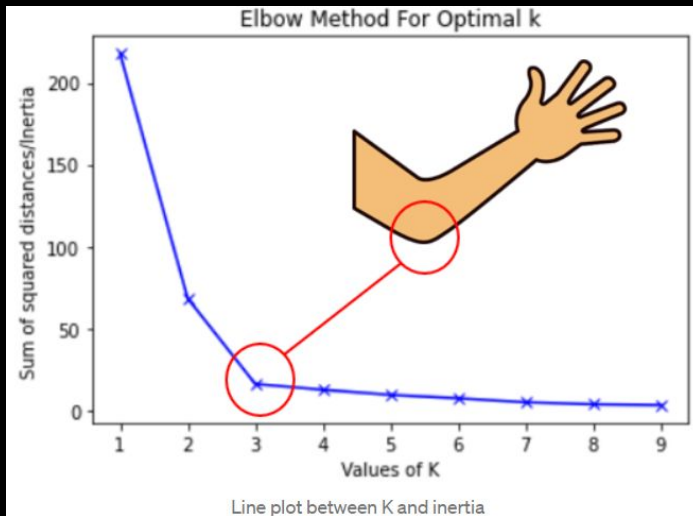
Por exemplo: se k for igual a 3. Então o algoritmo vai agrupar os usuários em 3 grupos diferentes com base nas suas características:



Portanto dois fatores influenciam na acuracidade da segmentação dos clientes: 1- Características do usuário fornecidas como entrada e 2- Quantidade k de clusters selecionados

MÉTODO DO COTOVELO

É executado o K-Means para várias quantidades diferentes de clusters e dizer qual dessas quantidades é o número ótimo de clusters. Nessa análise foi considerado de 2 até 50 clusters.



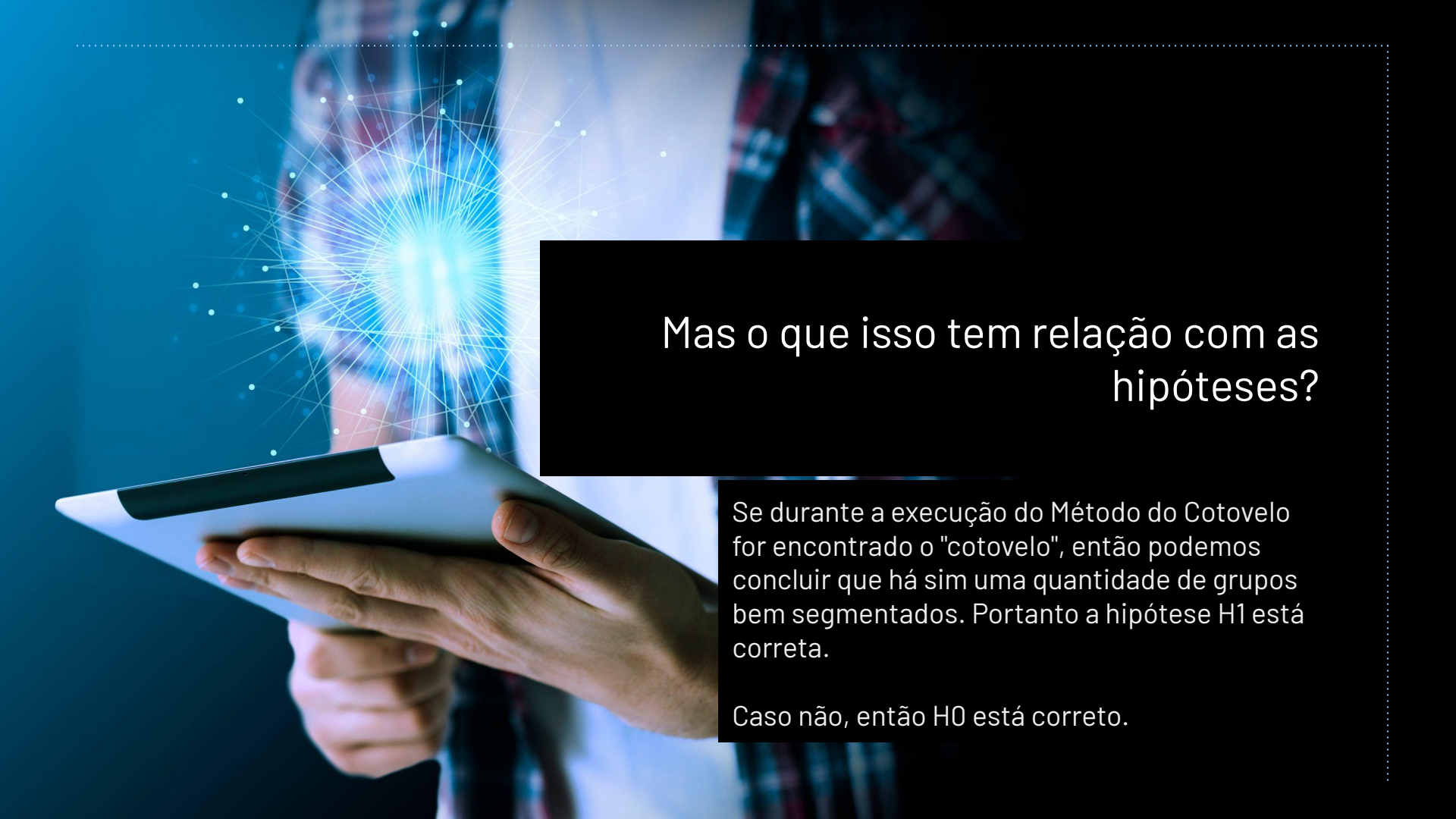
Exemplo do Método do Cotovelo - Fonte: Google

MÉTODO DO COTOVELO

O “Método do cotovelo” utiliza a distância média entre os pontos e centróide (ponto central do cluster), o ideal é que a distância seja a menor.



Exemplo do Método do Cotovelo – Fonte: Google

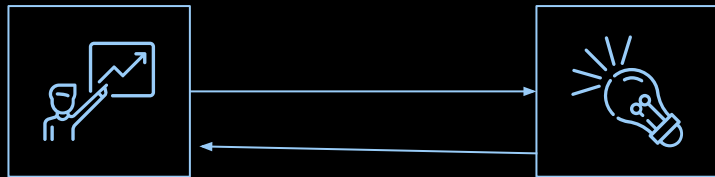
A person wearing a plaid shirt is holding a tablet. Overlaid on the image is a glowing blue network graphic with many nodes and connecting lines, resembling a data visualization or a neural network. The background is dark blue.

Mas o que isso tem relação com as hipóteses?

Se durante a execução do Método do Cotovelo for encontrado o "cotovelo", então podemos concluir que há sim uma quantidade de grupos bem segmentados. Portanto a hipótese H1 está correta.

Caso não, então H0 está correto.

ESTRATÉGIA UTILIZADA



SELEÇÃO DAS VARIÁVEIS

**RODAR O MÉTODO
DO COTOVELO (2
ATÉ 50 CLUSTERS)**



PARA IDENTIFICAR A MELHOR COMBINAÇÃO DE CARACTERÍSTICAS DO USUÁRIO, 4 TESTES FORAM FEITOS:

01

TESTE 1

Variáveis desconsideradas:
'userid','gender','address_state',
'perc_sesoes_click',
'perc_sesoes_clicks_1_tema','p
erc_sesoes_clicks_2_temas','p
erc_sesoes_clicks_3_mais_te
mas'

02

TESTE 2

Foi desconsiderado as
mesmas variáveis do
Teste 1, porém foi
incluído como entrada
pro modelo o estado
do usuário

03

TESTE 3

Variáveis
desconsideradas:
'userid','gender','addre
ss_state',
'perc_sesoes_click'

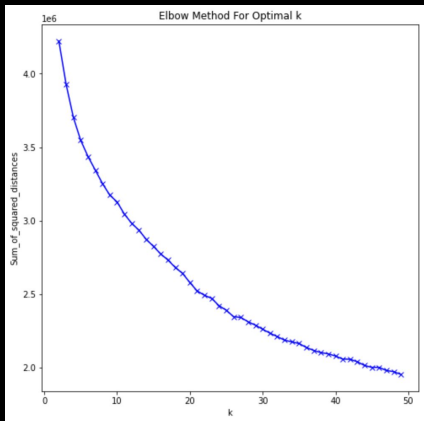
04

TESTE 4

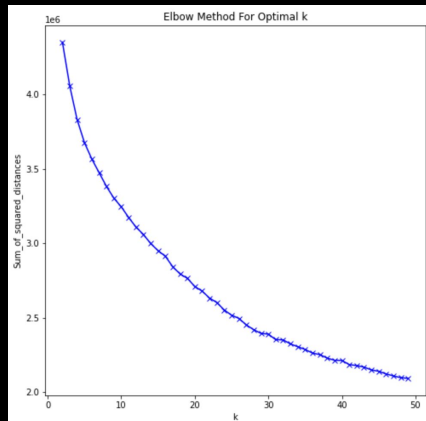
Foi desconsiderado as
mesmas variáveis do
Teste 3, além da
remoção das variáveis
com os nomes dos
temas



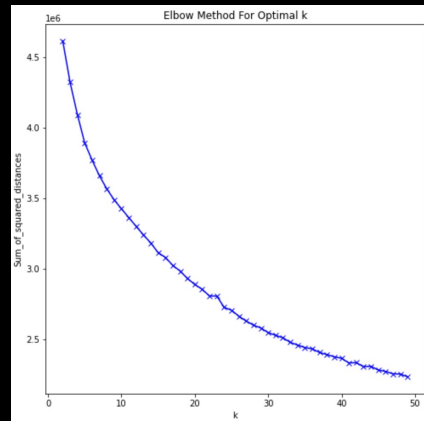
RESULTADOS DOS TESTES FEITOS



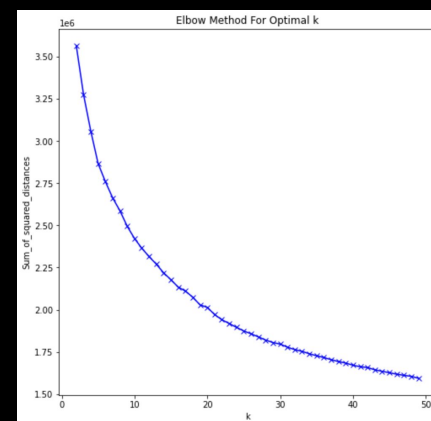
Teste 1



Teste 2



Teste 3



Teste 4

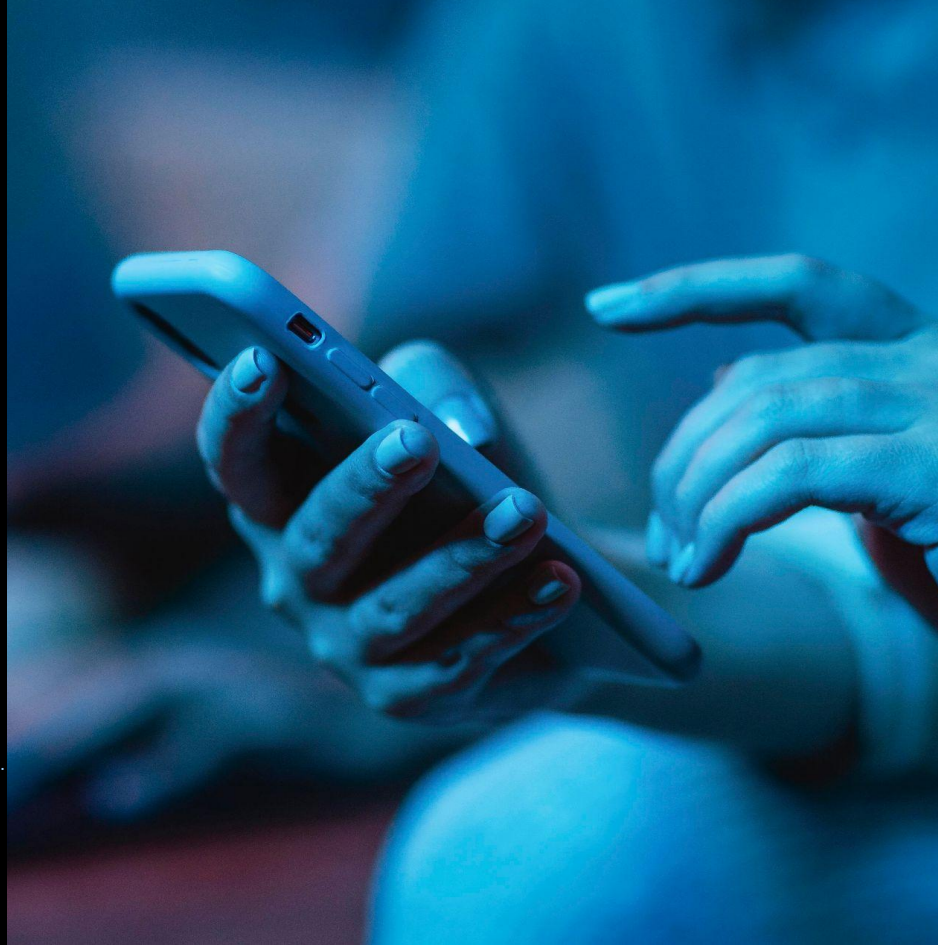
Podemos visualizar que há uma dificuldade em encontrar o valor K , pois a quebra não está bem definida. Portanto não tem como encontrar o "cotovelo"



CONCLUSÃO DAS HIPÓTESES

Com base nos testes que foram feitos para segmentação de usuários, não foi possível encontrar o melhor valor k (quantidade de clusters ideais).

Confirmando que a hipótese nula (H_0) aparenta ser a correta, ou seja, temos perfis mistos dentro da base de dados.





PRINCIPAIS INSIGHTS

PRINCIPAIS INSIGHTS

- Temas mais acessados pelos usuários: Jornalismo, Esporte e Entretenimento;
- Pessoas com interesse em jornalismo e entretenimento têm interesse em diversos temas;
- Pessoas com percentual de sessões no tema outros, tem interesse em conteúdos de vídeo;
- 75% dos usuários retornam ao menos 5 vezes na semana e mensalmente 17 vezes;
- Sessões de madrugada costumam ser pelo celular. E sessões de tarde costumam ser pelo computador;
- Mulheres têm uma distribuição maior de acesso a temas diversos
- Quanto mais a pessoa acessa pelo celular menos sessões ela tem pelo computador;
- Pessoas entre 30 e 49 anos têm mais sessões pelo celular. E pessoas acima de 70 anos tem menos sessões por esse dispositivo;
- Em média 33% das sessões dos usuários foram acessado na parte da tarde;
- Com base nos testes feitos com K-means foi concluído que a Hipótese H0: Existem grupos com interesse misto e a faz sentido ter um mix de tema.



DÚVIDAS?

Laura Damaceno de Almeida

Você tem alguma pergunta?

laura.da.almeida@hotmail.com

<https://lauradaalmeida.wixsite.com/laura-damaceno-portf>

