ll,9

**Controle do Documento**

**Histórico de revisões**

| **Data** | **Autor** | **Versão** | **Resumo da atividade** |
| --- | --- | --- | --- |
| 08/08/2022 | Marcos, Priscila, Maria Luísa, Matheus e Henrique | 1.1 | Criação do documento  Edição do tópico 4.1.1  Edição do tópico 4.1.4  Edição do tópico 4.1.5 |
| 09/08/2022 | Pedro  Priscila  Marcos | 1.2 | Edição do tópico 4.1.2  Edição do tópico 4.1.3  Inserção de dados nos subtópicos do tópico 4 |
| 10/08/2022 | Maria Luísa  Marcos  Pedro | 1.3 | Revisão e conclusão dos tópicos do artefato 1 (4.1.1, 4.1.2, 4.1.3, 4.1.4 e 4.1.5) |

**Sumário**

[**1. Introdução**](#_heading=h.2et92p0) **[5](#_heading=h.2et92p0)**

[**2. Objetivos e Justificativa**](#_heading=h.tyjcwt) **[6](#_heading=h.tyjcwt)**

[2.1. Objetivos](#_heading=h.3dy6vkm) **[6](#_heading=h.1t3h5sf)**

[2.2. Justificativa](#_heading=h.4d34og8) [6](#_heading=h.4d34og8)

[**3. Metodologia**](#_heading=h.2s8eyo1) **[7](#_heading=h.2s8eyo1)**

[3.1. CRISP-DM](#_heading=h.17dp8vu) [7](#_heading=h.17dp8vu)

[3.2. Ferramentas](#_heading=h.3rdcrjn) [7](#_heading=h.3rdcrjn)

[3.3. Principais técnicas empregadas](#_heading=h.26in1rg) [7](#_heading=h.26in1rg)

[**4. Desenvolvimento e Resultados**](#_heading=h.lnxbz9) **[8](#_heading=h.lnxbz9)**

[4.1. Compreensão do Problema](#_heading=h.35nkun2) [8](#_heading=h.35nkun2)

[4.1.1. Contexto da indústria](#_heading=h.1ksv4uv) [8](#_heading=h.1ksv4uv)

[4.1.2. Análise SWOT](#_heading=h.44sinio) [8](#_heading=h.44sinio)

[4.1.3. Planejamento Geral da Solução](#_heading=h.2jxsxqh) [8](#_heading=h.2jxsxqh)

[4.1.4. Value Proposition Canvas](#_heading=h.z337ya) [8](#_heading=h.z337ya)

[4.1.5. Matriz de Riscos](#_heading=h.3j2qqm3) [8](#_heading=h.3j2qqm3)

[4.1.6. Personas](#_heading=h.1y810tw) [9](#_heading=h.1y810tw)

[4.1.7. Jornadas do Usuário](#_heading=h.4i7ojhp) [9](#_heading=h.4i7ojhp)

[4.2. Compreensão dos Dados](#_heading=h.2xcytpi) [10](#_heading=h.2xcytpi)

[4.3. Preparação dos Dados](#_heading=h.1ci93xb) [11](#_heading=h.1ci93xb)

[4.4. Modelagem](#_heading=h.3whwml4) [12](#_heading=h.3whwml4)

[4.5. Avaliação](#_heading=h.qsh70q) [13](#_heading=h.qsh70q)

[4.6 Comparação de Modelos](#_heading=h.be1cqj72p9wo) [14](#_heading=h.be1cqj72p9wo)

[**5. Conclusões e Recomendações**](#_heading=h.3as4poj) **[14](#_heading=h.3as4poj)**

[**6. Referências**](#_heading=h.1pxezwc) **[15](#_heading=h.1pxezwc)**

[**Anexos**](#_heading=h.49x2ik5) **16**

# 1. Introdução

*Apresente de forma sucinta o parceiro de negócio, seu porte, local, área de atuação e posicionamento no mercado. Maiores detalhes deverão ser descritos na seção 4*

*Descreva resumidamente o problema a ser resolvido (sem ainda mencionar a solução).*

*Caso utilize citações ao longo desse documento, consulte a norma ABNT NBR 10520. Sugerimos o uso do sistema autor-data para citações.*

A **Rede Gazeta de Comunicações**, ou **Rede Gazeta**, é um conjunto de mídia brasileiro localizado no estado do Espírito Santo. Possuindo mais de 500 funcionários, a Rede Gazeta é o maior grupo de comunicação do Espírito Santo, a empresa foi fundada em 1928, com o jornal A Gazeta, porém apenas no ano de 1976 a TV Gazeta surgiu, aproveitando o grande crescimento dos meios de comunicação em massa. Atualmente, eles contam com a presença de 16 veículos de comunicação, abrangendo a TV, rádio e internet. Contudo, nos últimos anos, viu-se necessária a criação de um meio para melhorar a média de audiência dessa emissora

Nos últimos anos essa emissora cresceu bastante pelas suas produções próprias, porém com a grande competitividade das plataformas de streaming e das redes sociais, é de plena importância a constante pesquisa e análise de dados para suas próximas empreitadas. Por isso, o grupo hefEStos realizou a criação de um software que através da análise de dados e a partir da inteligência artificial realiza um modelo preditivo que ajudará na previsão da audiência de novos programas.

O modelo de negócio básico utilizado nos meios de comunicação, os quais a Rede Gazeta atua são a criação de conteúdo de qualidade - gerando uma base de consumidores - e ao mesmo tempo a venda de publicidade a essas empresas, que desejam atingir seu público.

As tendências para o futuro desse mercado são a crescente digitalização dos canais. Um exemplo disso é que o próprio jornal A Gazeta, que desde o final do ano de 2020 se tornou totalmente digital e parou de ser veiculada na edição impressa.

# 2. Objetivos e Justificativa

## 2.1. Objetivos

*Descreva resumidamente os objetivos gerais e específicos do seu parceiro de negócios*

## 2.2. Justificativa

*Faça uma breve defesa de sua proposta de solução, escreva sobre seus potenciais, seus benefícios e como ela se diferencia.*

# 3. Metodologia

*Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.*

## 3.1. CRISP-DM

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo

## 3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Colaboratory)

## 3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

# 

# 4. Desenvolvimento e Resultados

## 4.1. Compreensão do Problema

### 4.1.1. Contexto da indústria

Com mais de 500 funcionários, a Rede Gazeta é o maior grupo de comunicação do Espírito Santo, a empresa foi fundada em 1928, com o jornal A Gazeta, porém apenas no ano de 1976 a TV Gazeta surgiu, aproveitando o grande crescimento dos meios de comunicação em massa. Atualmente, eles contam com a presença de 16 veículos de comunicação, abrangendo a TV, rádio e internet.

As principais concorrentes deste mercado de comunicação no estado do Espírito Santo são a TV Tribuna, afiliada do SBT, e a TV Vitória, afiliada da Record. Na grande maioria do tempo a TV Gazeta se apresenta mais forte que suas concorrentes, salvo algumas exceções ocasionadas por eventos específicos.

O único tipo de ameaça sofrido pela TV são os meios digitais, que contam com conteúdos “on demand”, ou seja, que o usuário consegue escolher o que vai assistir. Porém, agora comparando a Rede Gazeta com as outras emissoras, sua vantagem se estabelece pela confiança dos telespectadores, sendo ressaltada através dos dados.

Seus clientes, como os patrocinadores, antigamente não possuíam um grande poder de barganha, por não haver nenhum canal de comunicação tão abrangente quanto a TV, porém, com a evolução dos meios digitais, essa competição ficou mais acirrada, com outros locais para veiculação de anúncios publicitários, aumentando o poder de barganha a esses anunciantes, e assim, forçando aos detentores do meio a baixarem o preço.

Apesar de ser quase impossível a substituição da TV, por já ser um meio consolidado há várias décadas, é observada, cada vez mais, a transferência de certa parcela do seu público para os meios digitais.

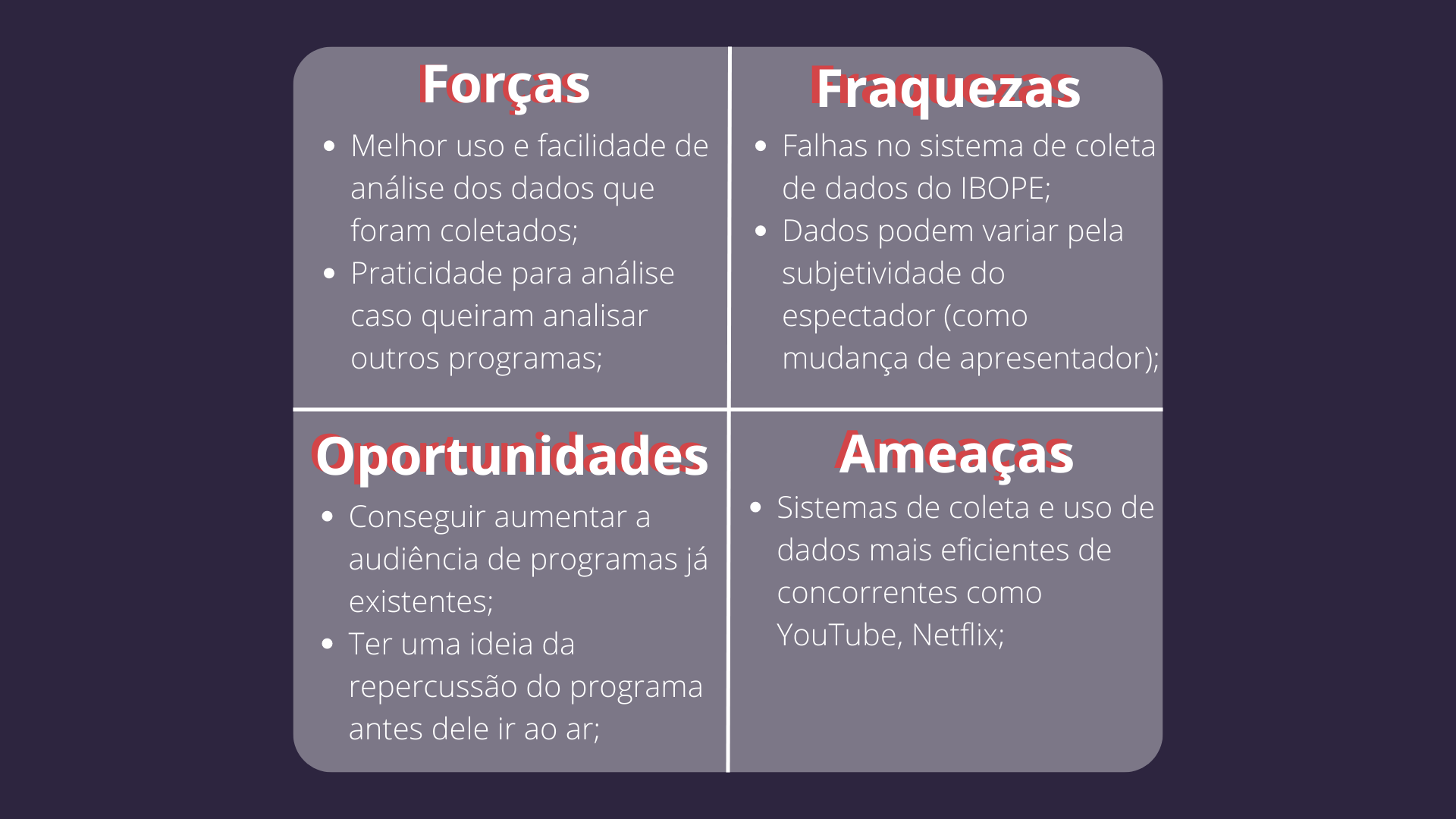
A necessidade de antecipar tendências e estar sempre atualizado com as novidades do mercado é o principal ponto em que a solução desenvolvida pelo grupo hefEStos atua mais fortemente. Conseguindo prever as audiências mais rapidamente e ficar a par das novidades do mercado, ajudando a identificar a reação do público na previsão do score de audiência de um novo programa.

### 

### 

### 4.1.2. Análise SWOT

Figura 1 - Matriz SWOT



Fonte: do próprio autor.

### 4.1.3. Planejamento Geral da Solução

*a) quais os dados disponíveis (fonte e conteúdo - exemplo: dados da área de Compras da empresa descrevendo seus fornecedores)*

Para o desenvolvimento da aplicação a Rede Gazeta disponibilizou alguns dados referentes à audiência, são esses:

* Datas;
* Hora de início;
* Emissoras;
* Dias da semana;
* Porcentagens utilizadas:
  + Rat%
  + Shr%
  + Rch%
  + Fid%
* Total de domicílios;
* Caracterização do público telespectador:
  + Classe:
    - AB;
    - C1;
    - C2;
    - DE;
  + Gênero:
    - Masculino;
    - Feminino;
  + Faixa etária:
    - 4-11 anos;
    - 12-17 anos;
    - 18-24 anos;
    - 25-34 anos;
    - 35-49 anos;
    - 50-59 anos;
    - 60+ anos.
* Grade de programação de cada emissora;

Divididos nas seguintes emissoras:

* Emissora 1;
* Emissora 2;
* Emissora 3;
* Canais pagos;
* Total Ligados Especial (TLE);
* Não identificado (NI);

Além disso, é dividido em dias da semana:

* Segunda a sexta;
* Sábado;
* Domingo.

*b) qual a solução proposta (pode ser um resumo do texto da seção 2.2)*

Sabendo da necessidade do parceiro por uma análise preditiva da audiência da Rede Gazeta, a solução proposta busca prover por meio de uma Inteligência Artificial, descrever a pontuação que determinado programa terá, quando lançado, de acordo com o horário, dia da semana, eixo e público especificado.

*c) qual o tipo de tarefa (regressão ou classificação)*

O tipo de método que será empregado será o de *Regressão,* pois iremos estimar dados de audiência de acordo com os valores de entrada que foram coletados anteriormente.

*d) como a solução proposta deverá ser utilizada*

O usuário deverá inserir os horários, a data, o segmento do novo programa e as características do público e a solução irá retornar uma previsão dos pontos de audiência, junto do peso de cada variável no modelo preditivo.

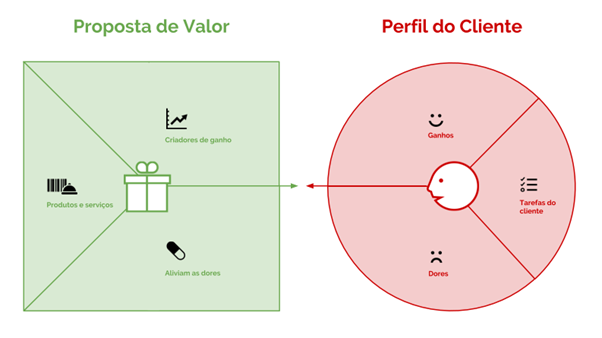
*e) quais os benefícios trazidos pela solução proposta*

Os benefícios trazidos pela solução são os de melhorar a capacidade na realização da análise dos dados, de ajudar em previsões para lançamentos de programas futuros, além de promover uma melhora no desempenho de programas já existentes.

*f) qual será o critério de sucesso e qual medida será utilizada para o avaliar*

O critério utilizado como parâmetro de sucesso será a comparação com os valores fornecidos pelo banco de dados. Nesse, utilizamos parte do banco de dados para o treinamento do modelo, e o restante foi utilizado para comprovar o quão próximo foi o retorno do modelo com relação à realidade.

### 4.1.4. Value Proposition Canvas



**--> Perfil do cliente**

**Tarefas do cliente:**

-Criação de conteúdo para programas de TV.

- Inserir os dados solicitados pelo software.

**Dores:**

- Dificuldade em determinar o conteúdo adequado para agradar a audiência em um

determinado horário.

-Alto investimento em programas que não repercutiram da forma esperada.

**Ganhos do cliente:**

-Melhora na acurácia da capacidade de prever a audiência de um determinado programa,

baseado em algumas de suas informações como horário, data e tema. Adaptando, assim,

o conteúdo, para aumentar o score de audiência.

**-->Mapa de Valor**

**Produtos e serviços:**

-Um modelo preditivo que recebe informações básicas de um possível novo programa,

e retorna um score de audiência e as principais variáveis que pesaram nesse.

**Analgésicos/alívio das dores:**

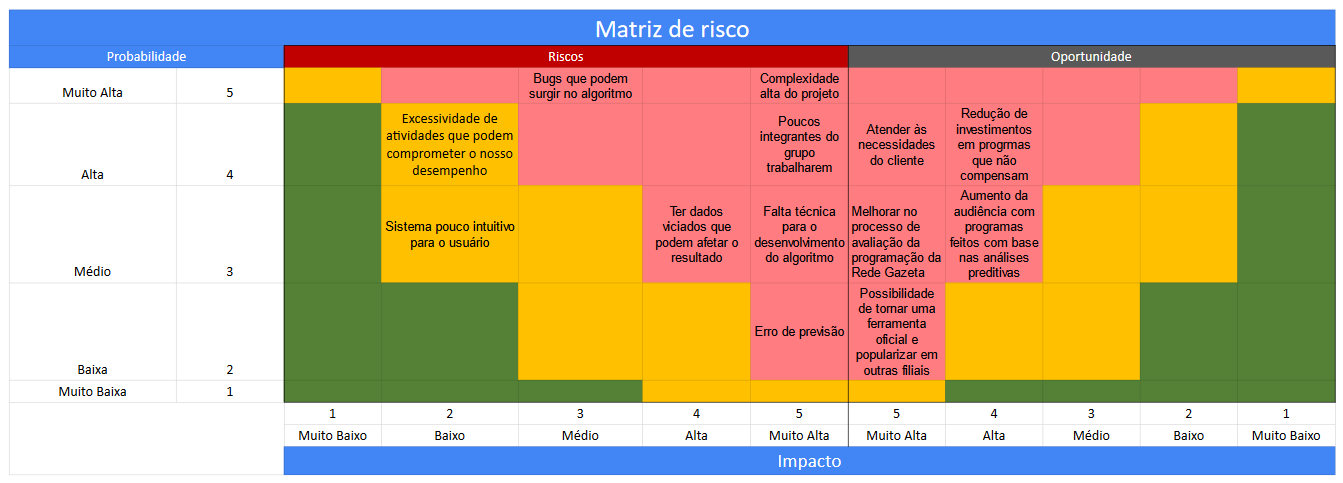
-Previsão acurada do possível sucesso de conteúdos, antes da produção.

**Criadores de ganhos:**

-Evita gastos com programas de baixa audiência.

-Fornece informações de conteúdos capazes de maximizar a audiência.

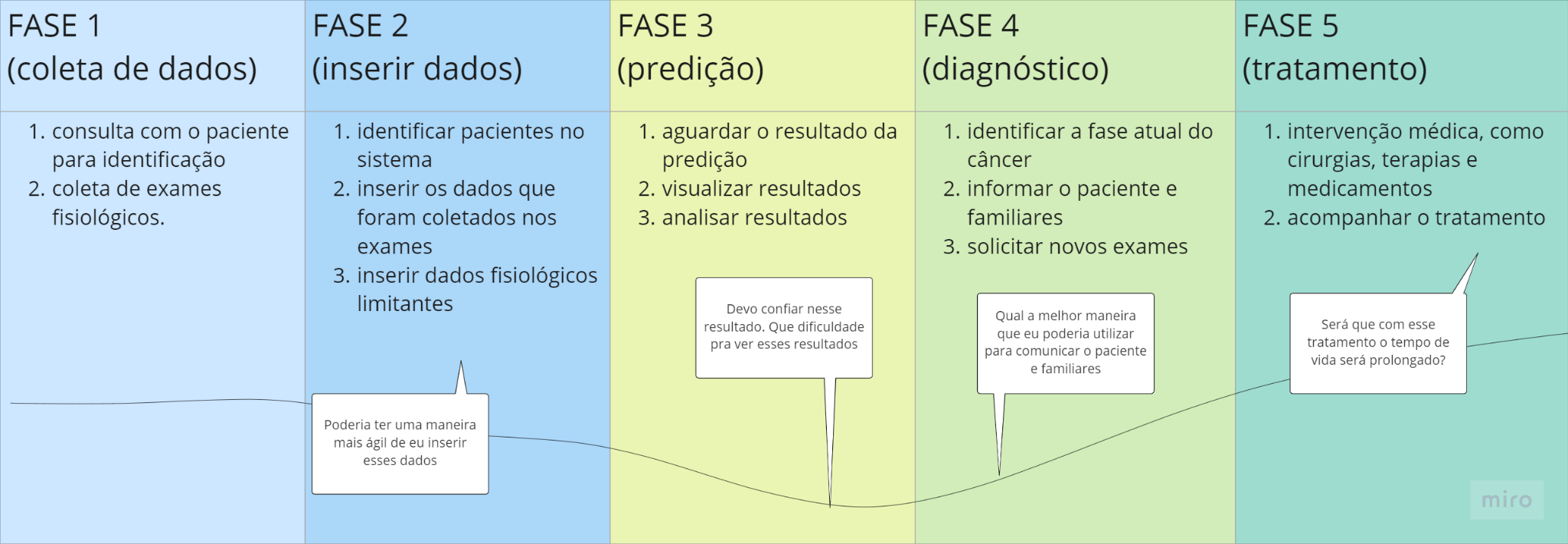
### 4.1.5. Matriz de Riscos



### 4.1.6. Personas

Posicione aqui suas Personas (as que utilizam o modelo e as que são afetadas pelo modelo)

### 4.1.7. Jornadas do Usuário



## 4.2. Compreensão dos Dados

1. *Descreva os dados a serem utilizados (disponibilizados pelo cliente e outros se tiverem sido incluídos), detalhando a fonte, o formato (CSV, XLSX, banco de dados, etc.), o conteúdo e o tamanho.*

Recebemos dados de diferentes emissoras no formato XLSX, que foi convertido posteriormente para CSV, baseados em pesquisas do IBOPE, utilizando parâmetros percentuais em: Rat%, essa medida é calculada a partir da quantidade de indivíduos/domicílios ligados em determinado evento de TV, sendo que 1 ponto de audiência se equivale a 1% do universo pesquisado. Esse é calculado da seguinte forma: Rat% = (Rat# / universo) x 100; Shr% descreve a participação da audiência em um determinado evento, sobre o total de televisores ligados, em um determinado período. Esse é calculado da seguinte maneira: Shr% = (Rat% / TLE%) x 100 ; Rch% é o total de indivíduos, ou domicílios, diferentes alcançados por pelo menos 1 minuto. Reforçando que o tempo total, nesse caso, não está sendo considerado, e sim o contato que houve com a programação, faixa horária, emissora, etc. Esse é calculado da seguinte forma: Rch% = número de telespectadores / universo x 100; Fid% ilustra a permanência dos telespectadores no evento em questão, ou seja, quanto tempo daquele programa foi consumido pelo público. Esse é calculado da seguinte maneira: Fid% = Rat% / Rch% x 100. Além disso, recebemos informações acerca do perfil da audiência, como gênero, faixa de idade e classe social.

* 1. *Se houver mais de um conjunto de dados, descrição de como serão agregados/mesclados.*

Devido ao alto volume de dados, para melhor visualização, esses foram mesclados usando a média dos valores, e agregados a partir de uma funcionalidade da plataforma google sheets. Ademais, na segunda semana do projeto, depois de múltiplos pedidos da turma, foi adicionado uma nova planilha no nosso conjunto de dados, que contempla a grade horária, com as seguintes informações: praça, data, faixa horária e o nome do programa e segmento das três emissoras que estamos trabalhando( Emissora 1, 2 e 3). Com essas novas informações, vamos poder ter uma noção de qual programa está fazendo o maior sucesso, comparando com as outras emissoras, e assim analisar o porquê isso acontece, se é por conta do horário ou pela falta de concorrência, etc.

* 1. Descrição dos riscos e contingências relacionados a esses dados (qualidade, cobertura/diversidade e acesso).

Nesse contexto, durante a análise dos dados, foi verificado que há dois riscos a serem considerados na análise de tais dados: o primeiro risco é os dados tenham viés, já o segundo se trata da concorrência injusta na coleta dos dados.

No primeiro ponto é necessário ponderar que, no aparelho utilizado na medição da audiência, quando colocado no domicílio de cada família, é criado um perfil para cada integrante, com informações de classe social, gênero e idade. Quando a TV é ligada, quem está assistindo deve selecionar o seu próprio perfil, é nesse momento que pode acontecer o viés, já que uma criança, por exemplo, pode selecionar errado ou mais de uma pessoa pode estar vendo TV. Adicionalmente, a TV pode continuar ligada em um perfil originalmente correto, mas que já não é válido, ou seja, outro integrante pode ter começado a ver TV.prestando atenção e não foi selecionado, ou até mesmo uma mãe/pai pode, na pressa, selecionar o perfil próprio para uma criança assistir.

O segundo risco é por conta da concorrência injusta, já que a emissora chamada de Não Identificado, abrange algo muito genérico, isto é, apesar de ser chamada de uma emissora, dentro dela ela é dividida por diversas plataformas de vídeos, tornando a comparação injusta para as emissoras 1,2 e 3, que estão separadas.

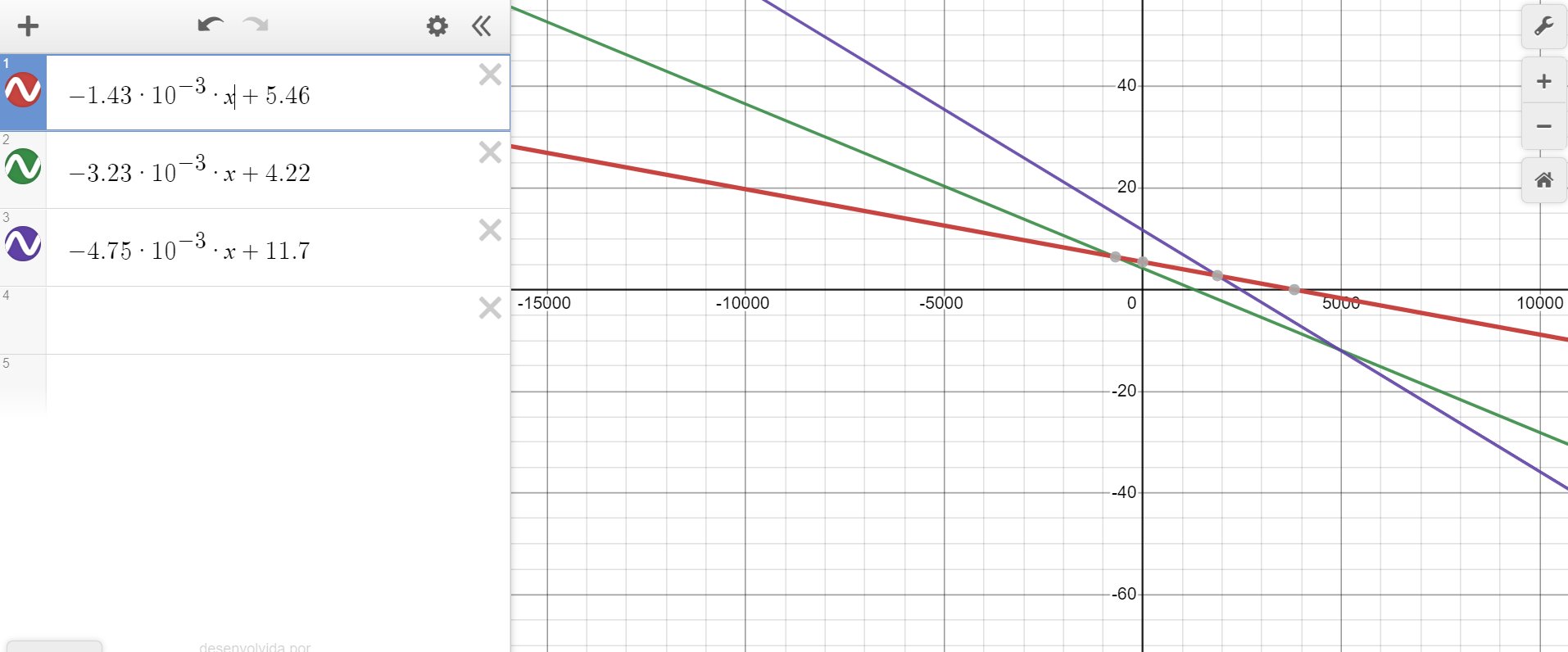
* 1. Se aplicável: descrição de como será selecionado o subconjunto para análises iniciais (quando o tamanho do conjunto de dados impossibilita a utilização do conjunto completo em todas as etapas da definição do modelo a ser usado).

—------------

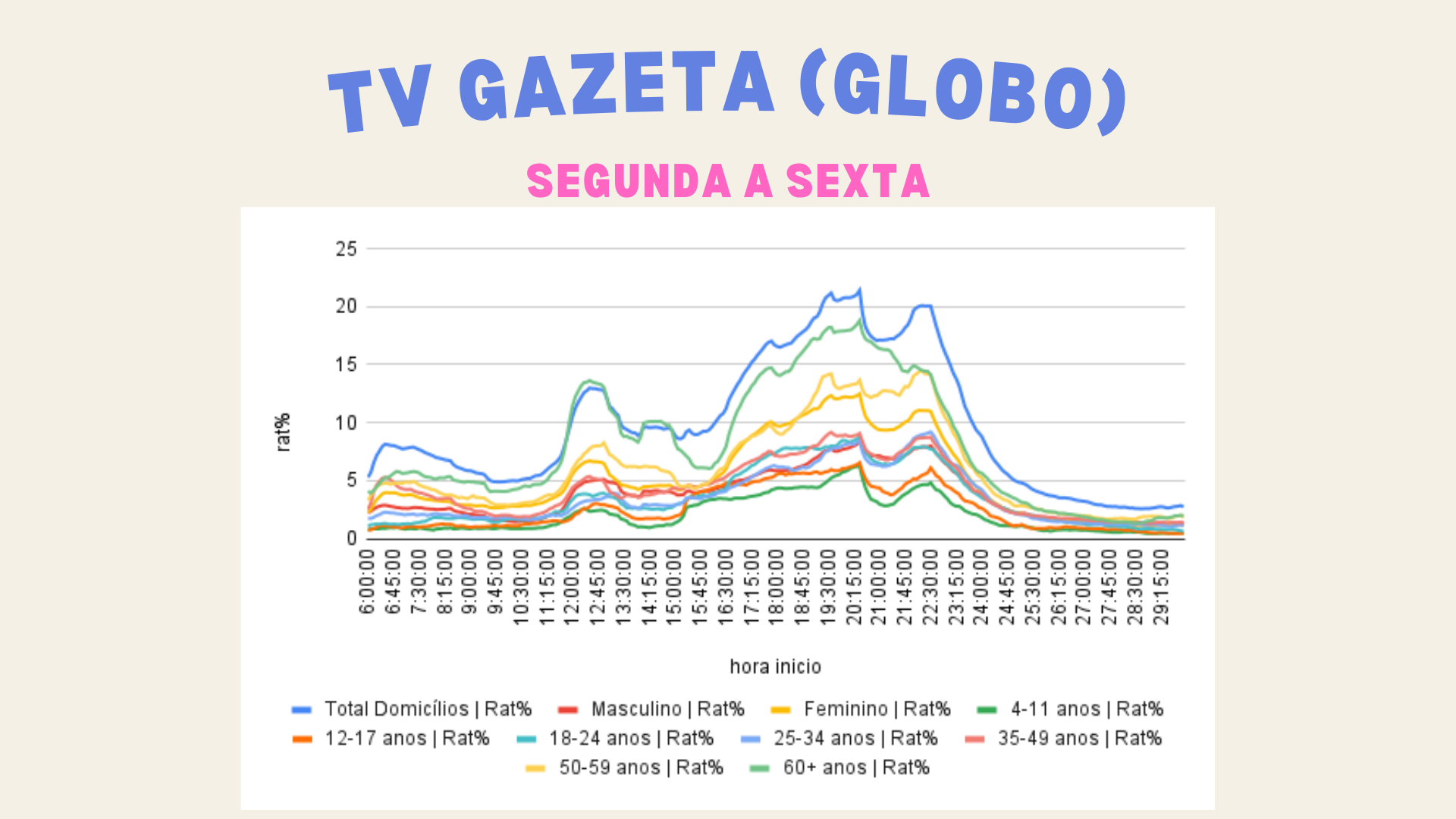
* 1. Se houver: descrição das restrições de segurança.

Para a elaboração do modelo, embora tenha sido concedido os dados de programação e audiência de algumas emissoras de TV aberta do Espírito Santo para a criação do modelo, foi proibida a publicação dos nomes das emissoras.

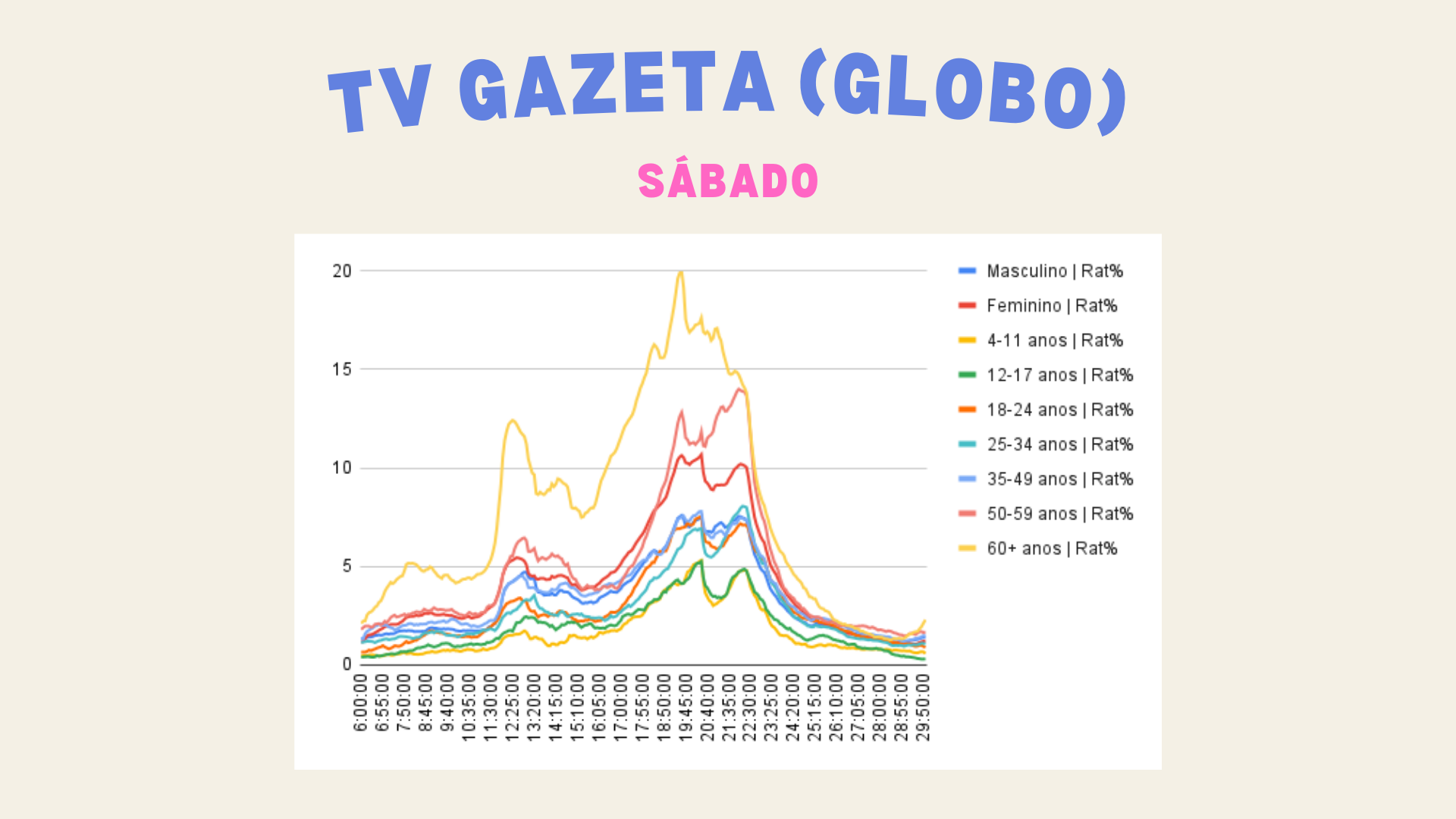
1. Descrição estatística básica dos dados, principalmente dos atributos de interesse, com inclusão de visualizações gráficas e como essas análises embasam suas hipóteses

Este gráfico ilustra a tendência na queda de audiência das 3 principais emissoras de televisão aberta nos últimos dois anos.

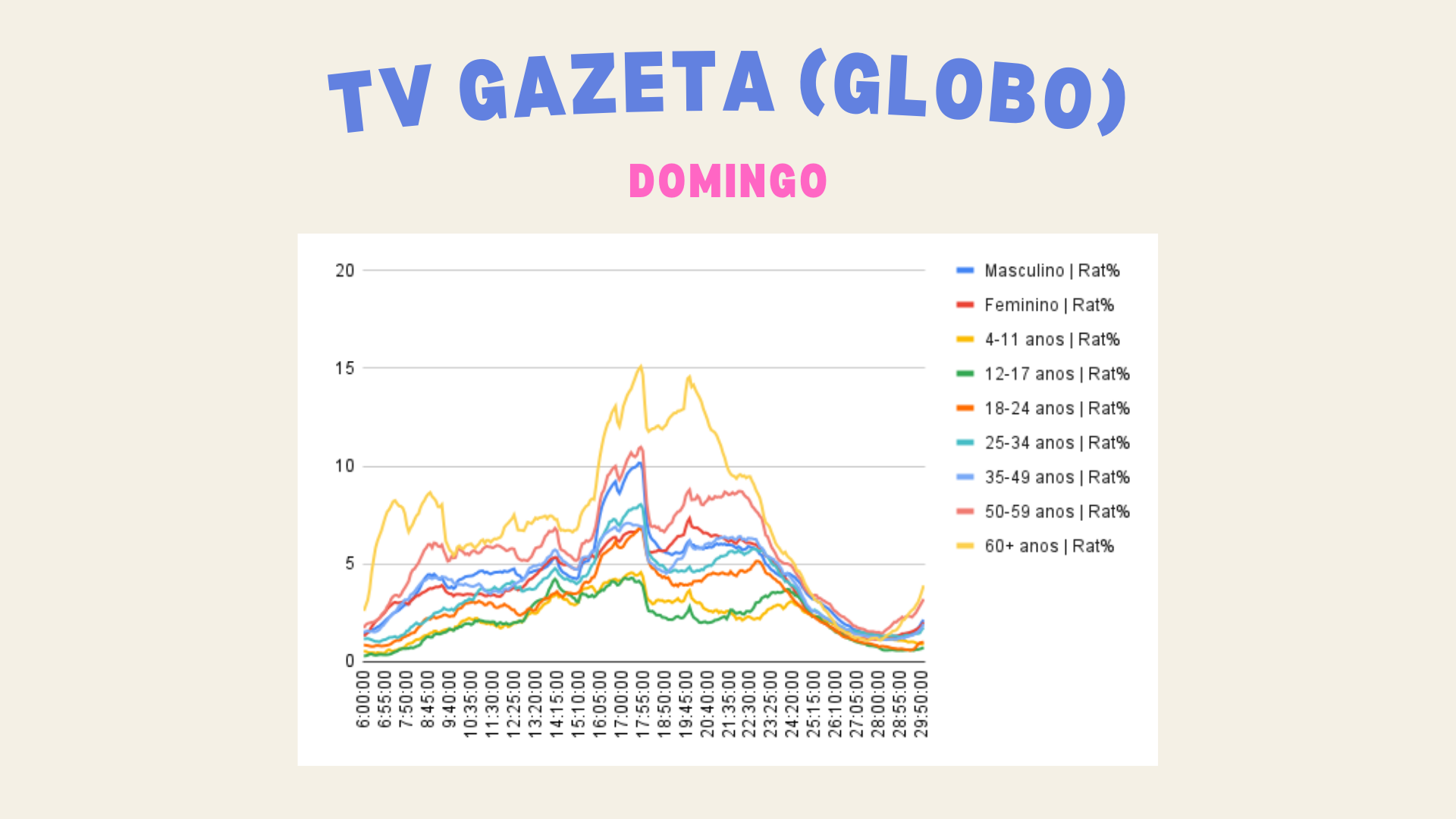
**EMISSORA 1:**



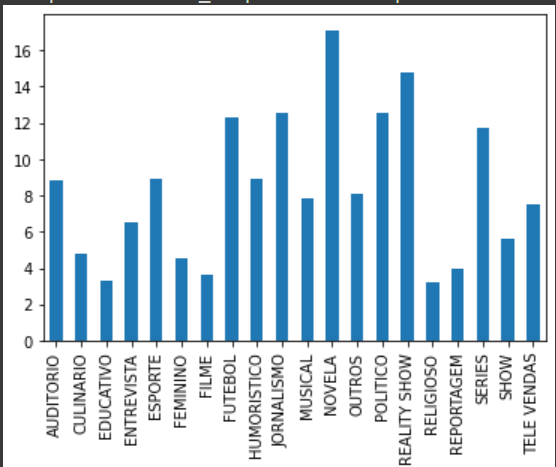
Este gráfico ilustra de forma nítida os dois grandes picos de audiência. Um ocorre no horário do almoço(aproximadamente às 12h), no momento em que está passando um programa jornalístico. Já o outro ocorre durante a noite (das 18h até aproximadamente 00h), que é o horário em que são exibidos programas de entretenimento como novelas e reality shows. Pode-se notar também que no horário de exibição de um jornal a noite(aproximadamente das 20h30 às 21h30), há uma redução do público.



No sábado a tendência de dois picos se mantém, porém a audiência se mantém alta durante a tarde.



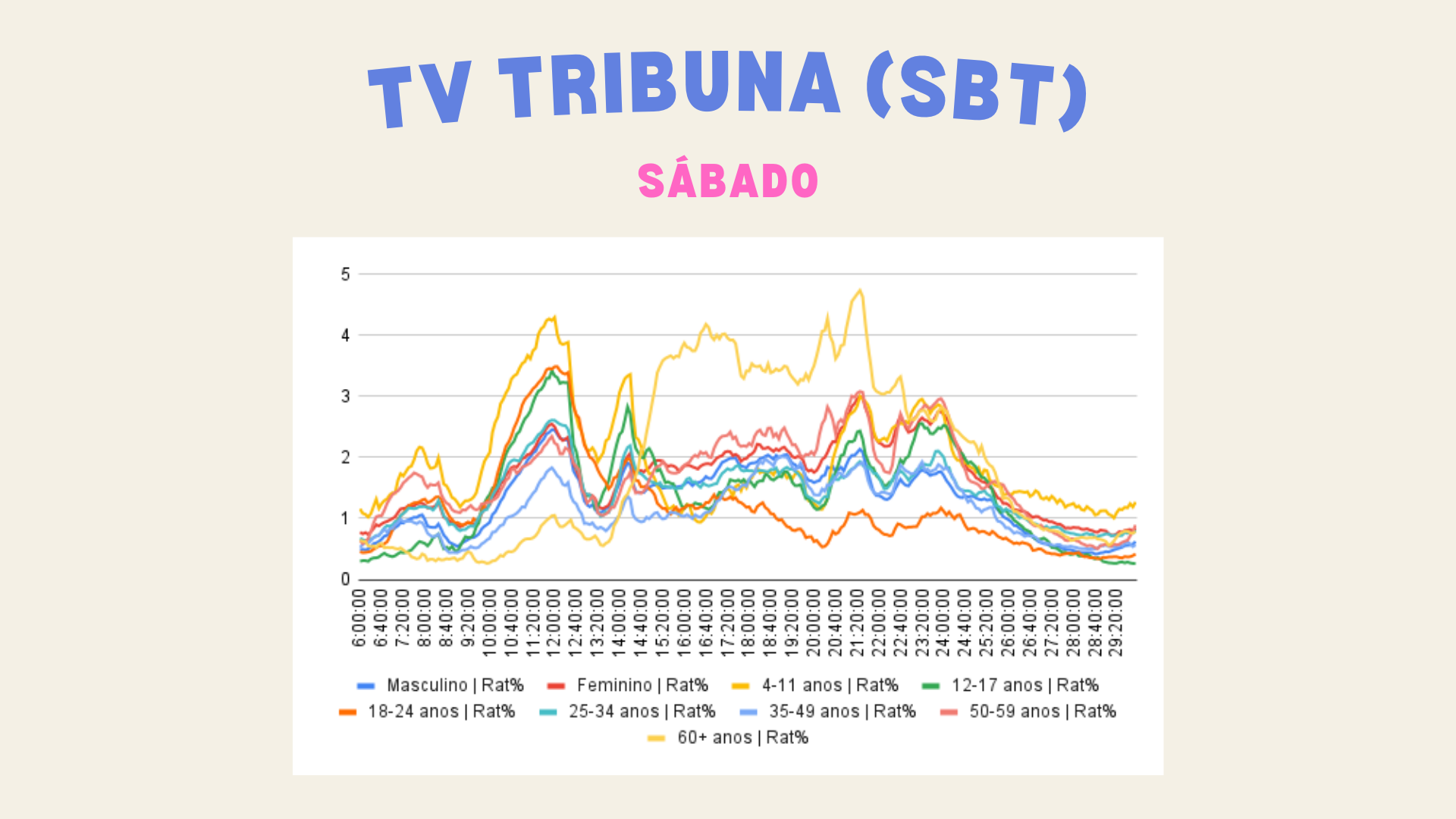
O domingo é um dia totalmente atípico, no qual a audiência possui menor discrepância durante o dia, mantendo uma boa audiência durante quase todo o tempo.



**EMISSORA 2:**

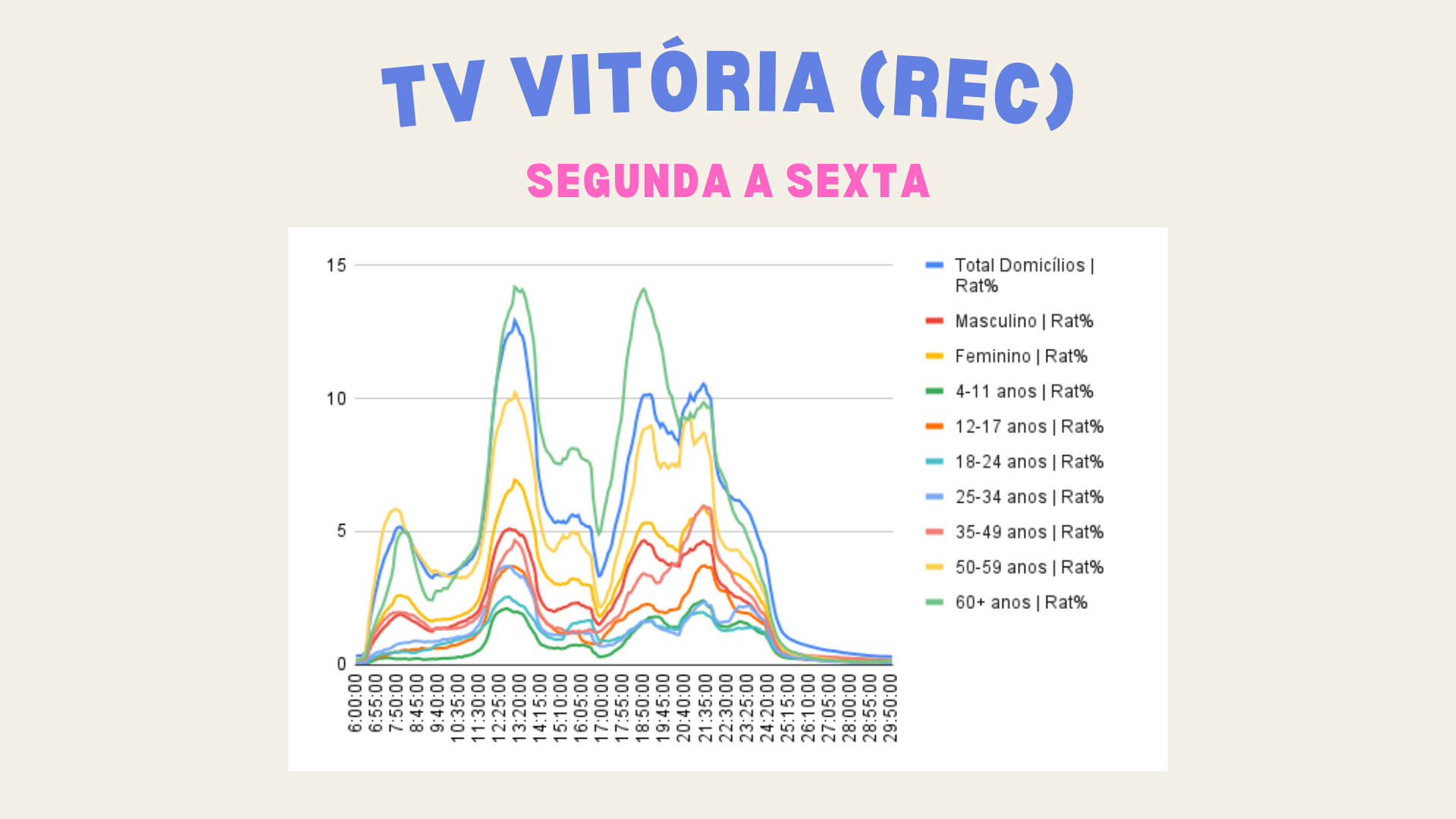


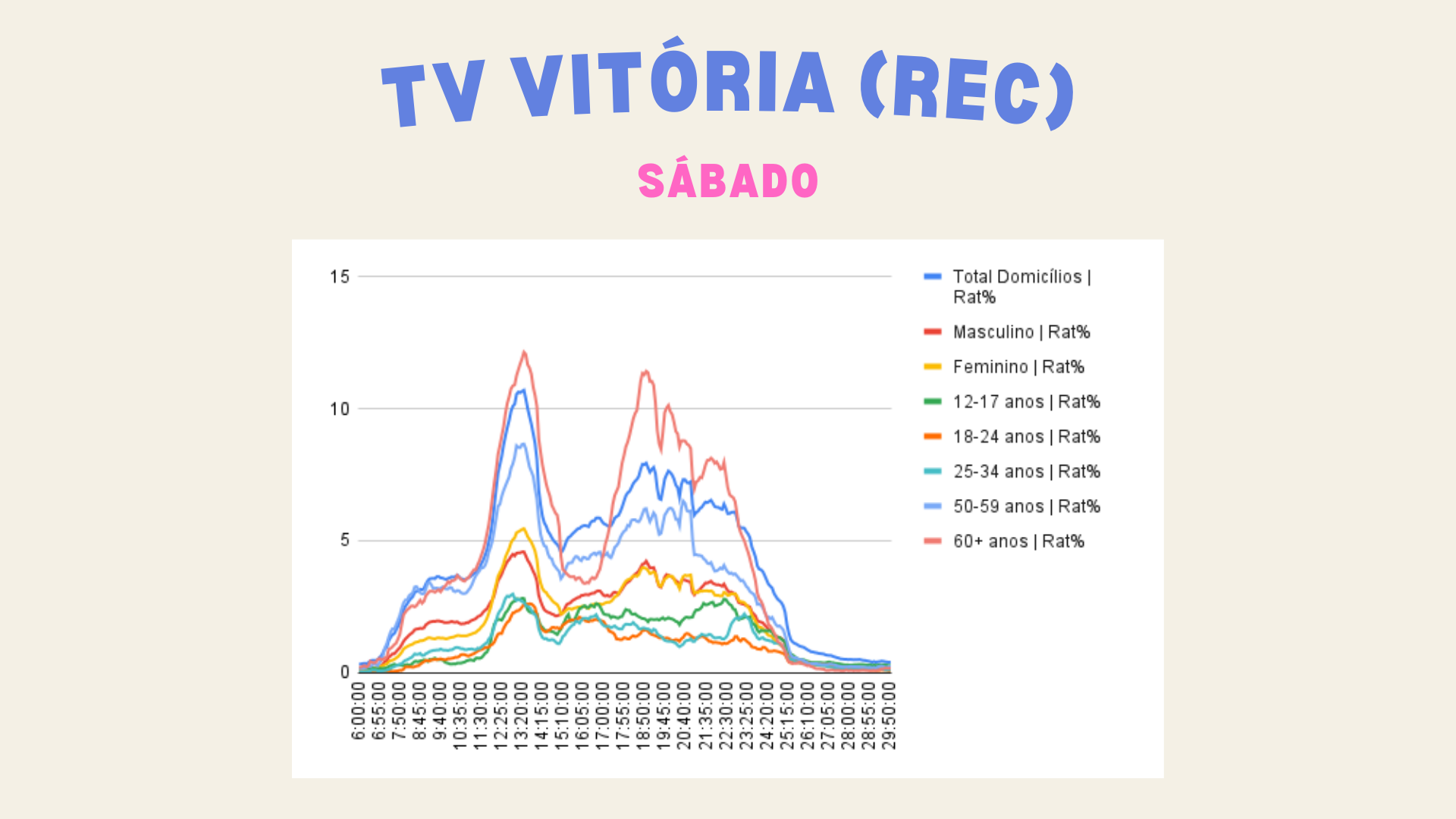
São notados 3 grandes audiências que ocorrem durante o almoço, às 19 e às 21h.

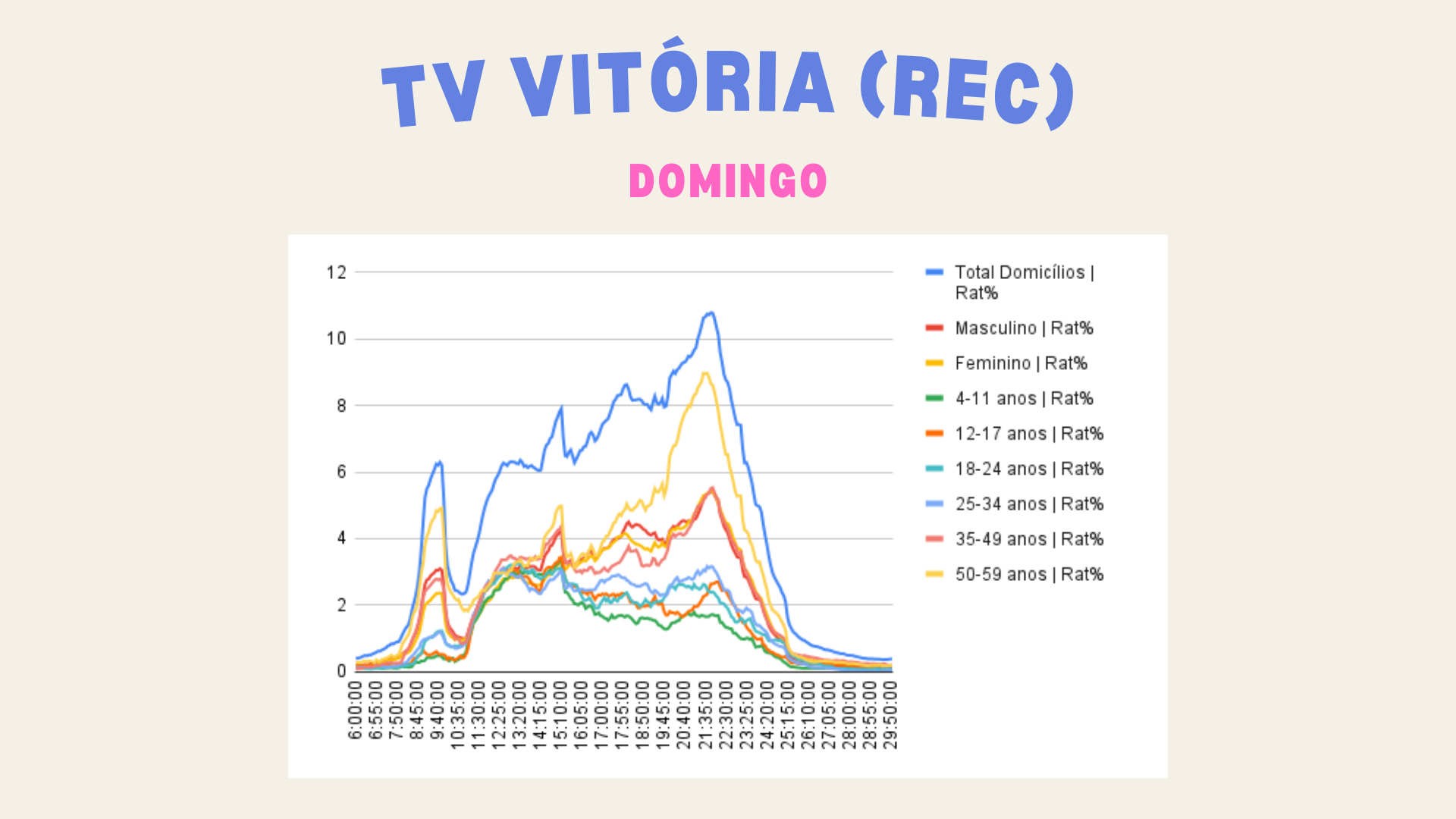




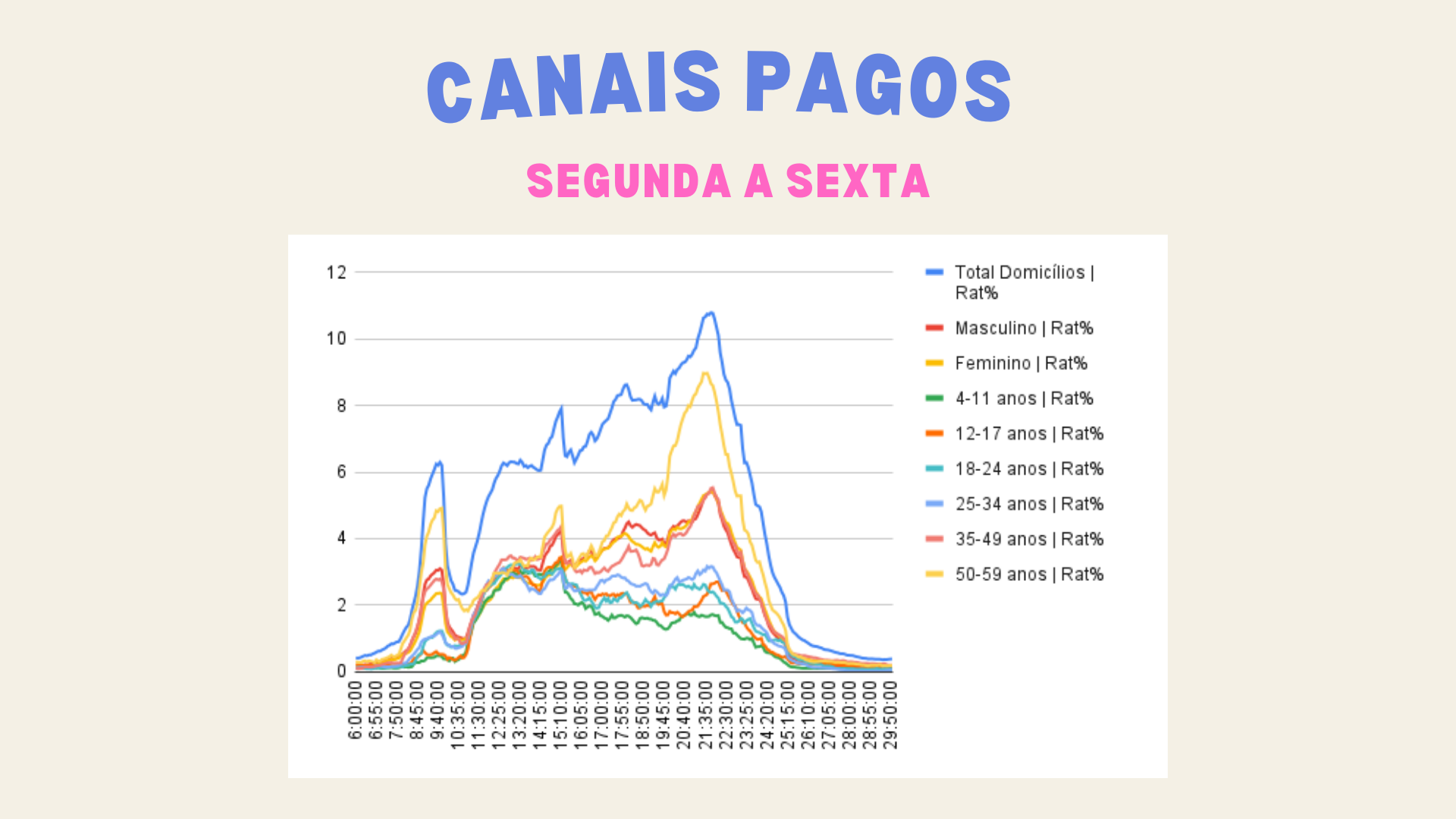
**EMISSORA 3:**

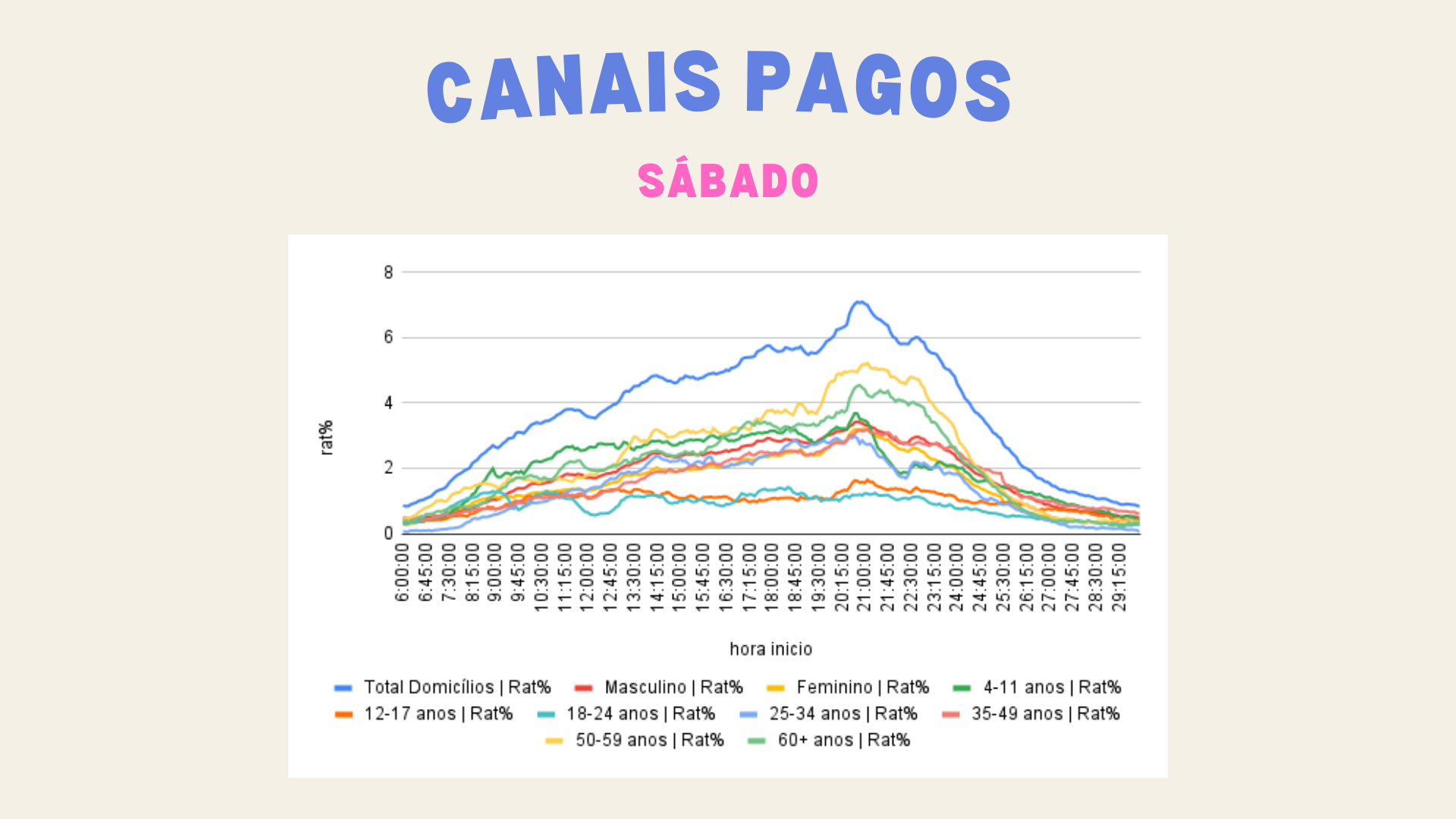
****

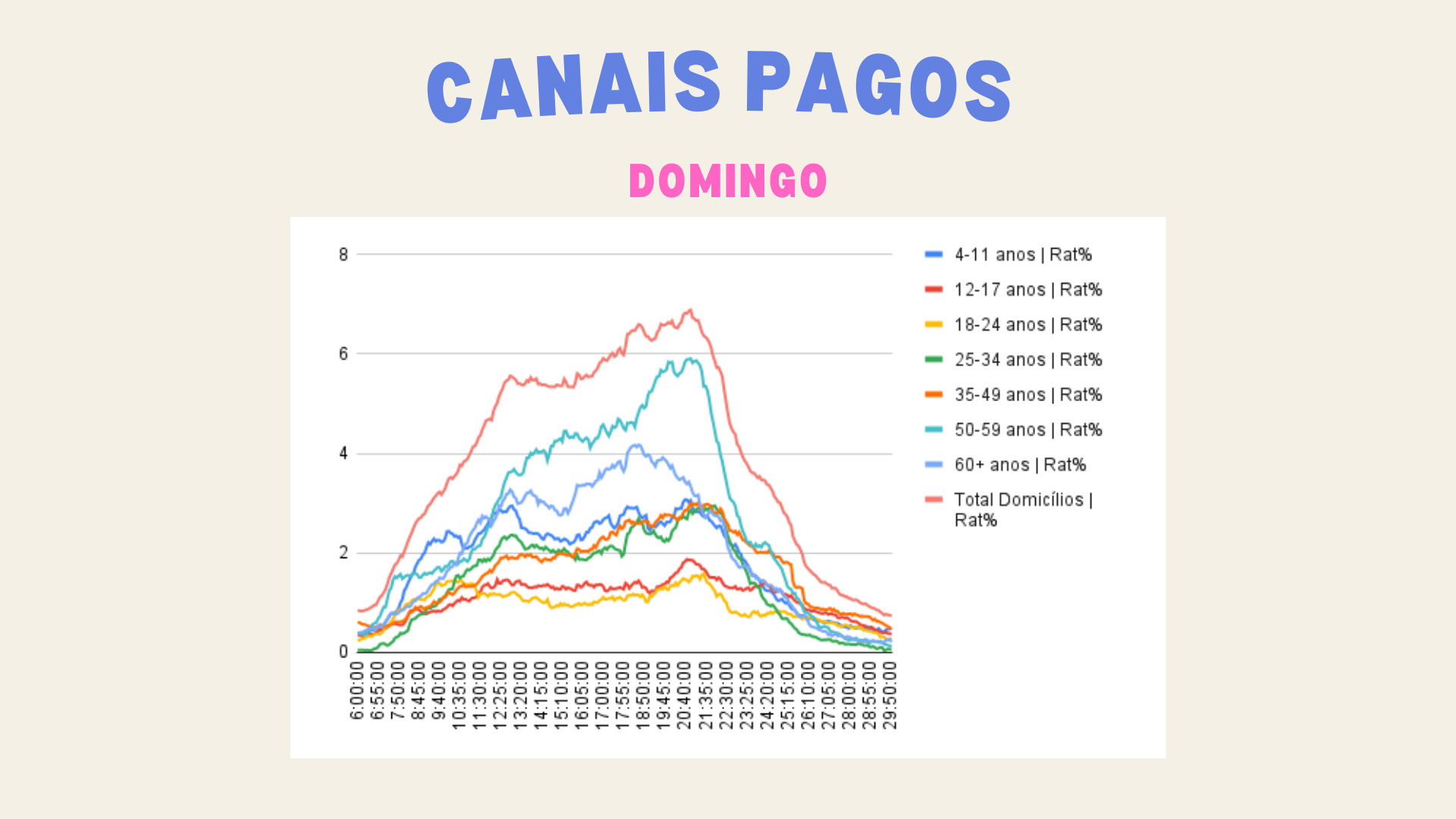
****

****

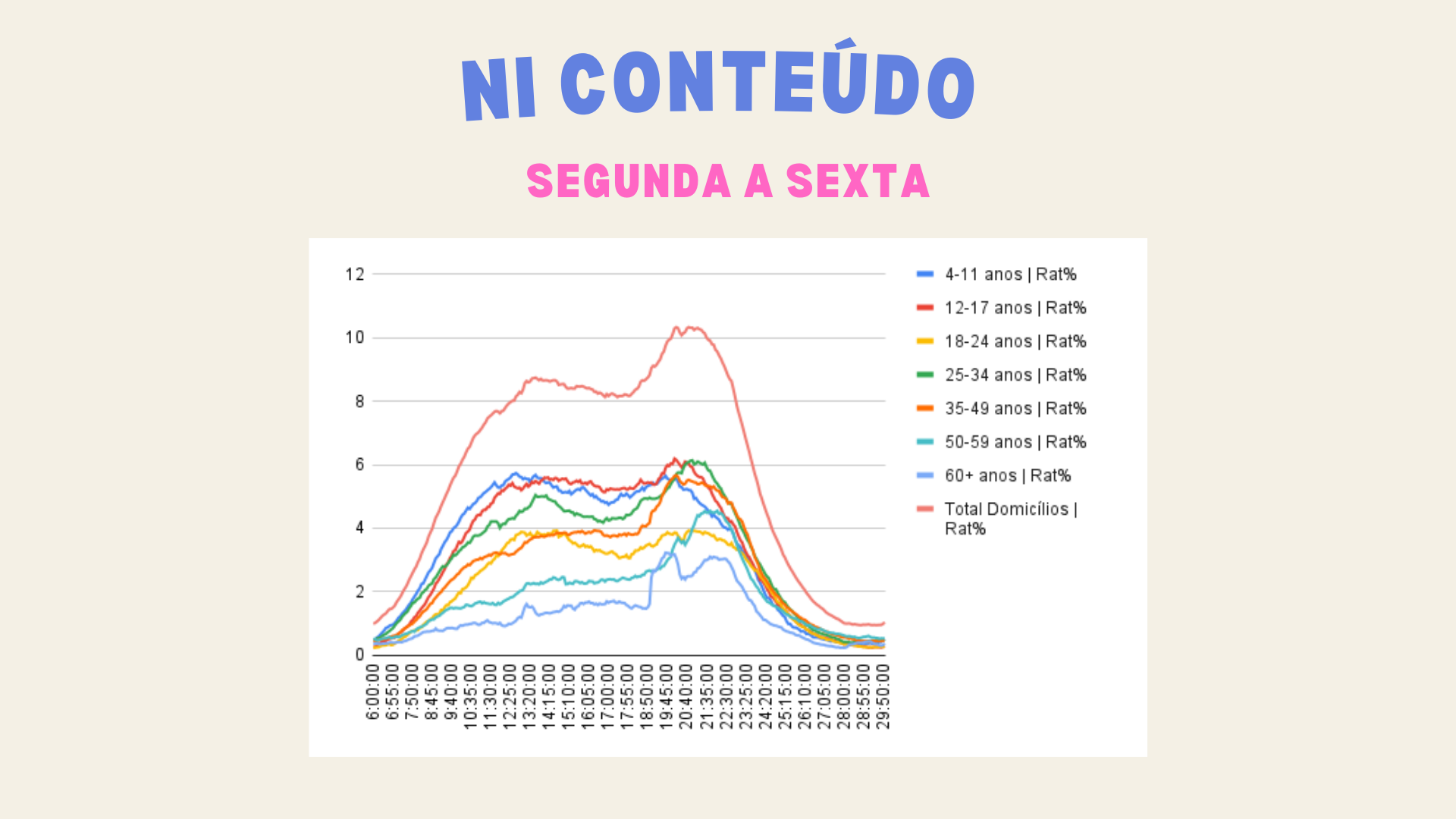
**CANAIS PAGOS:**

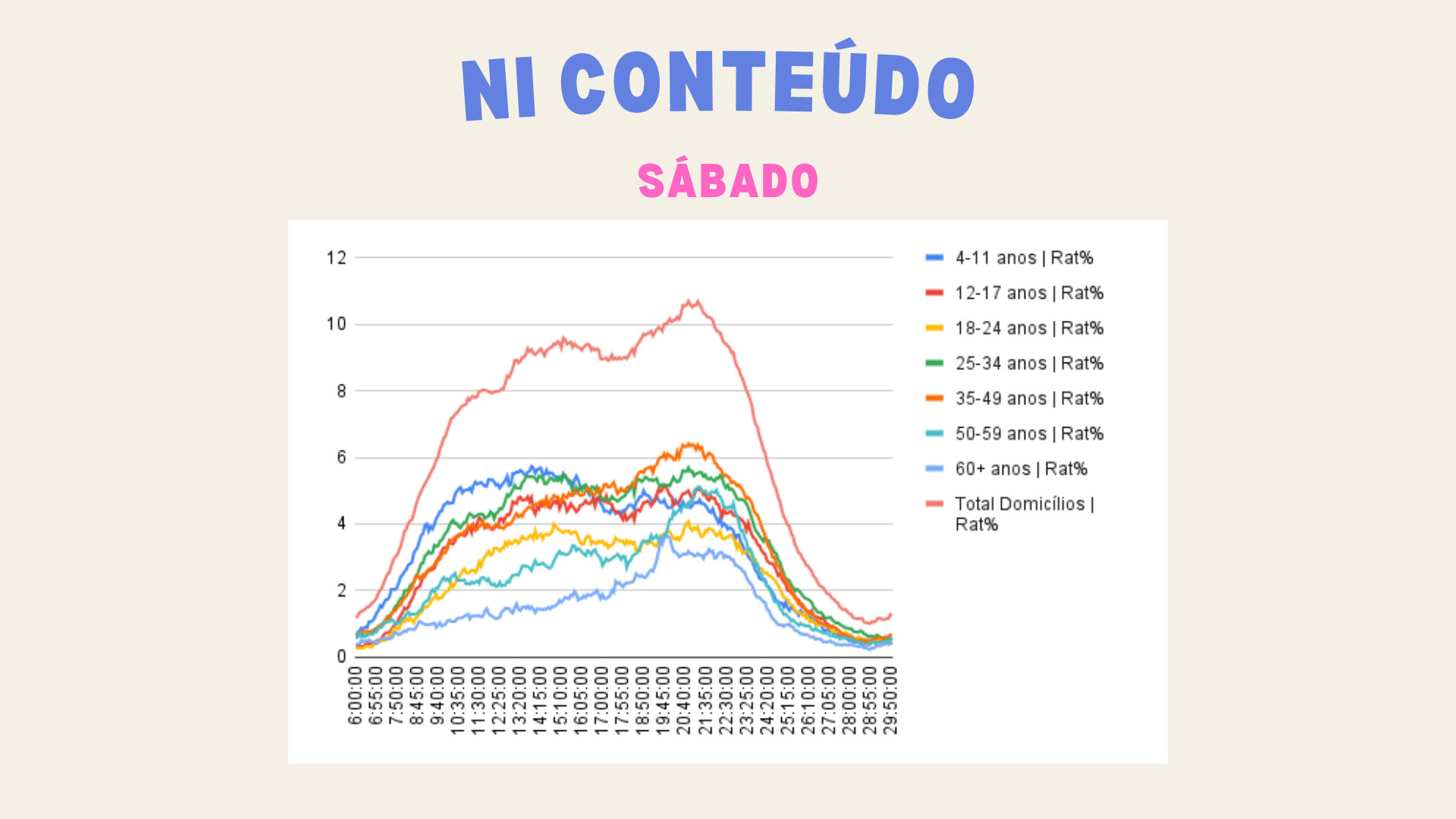
****

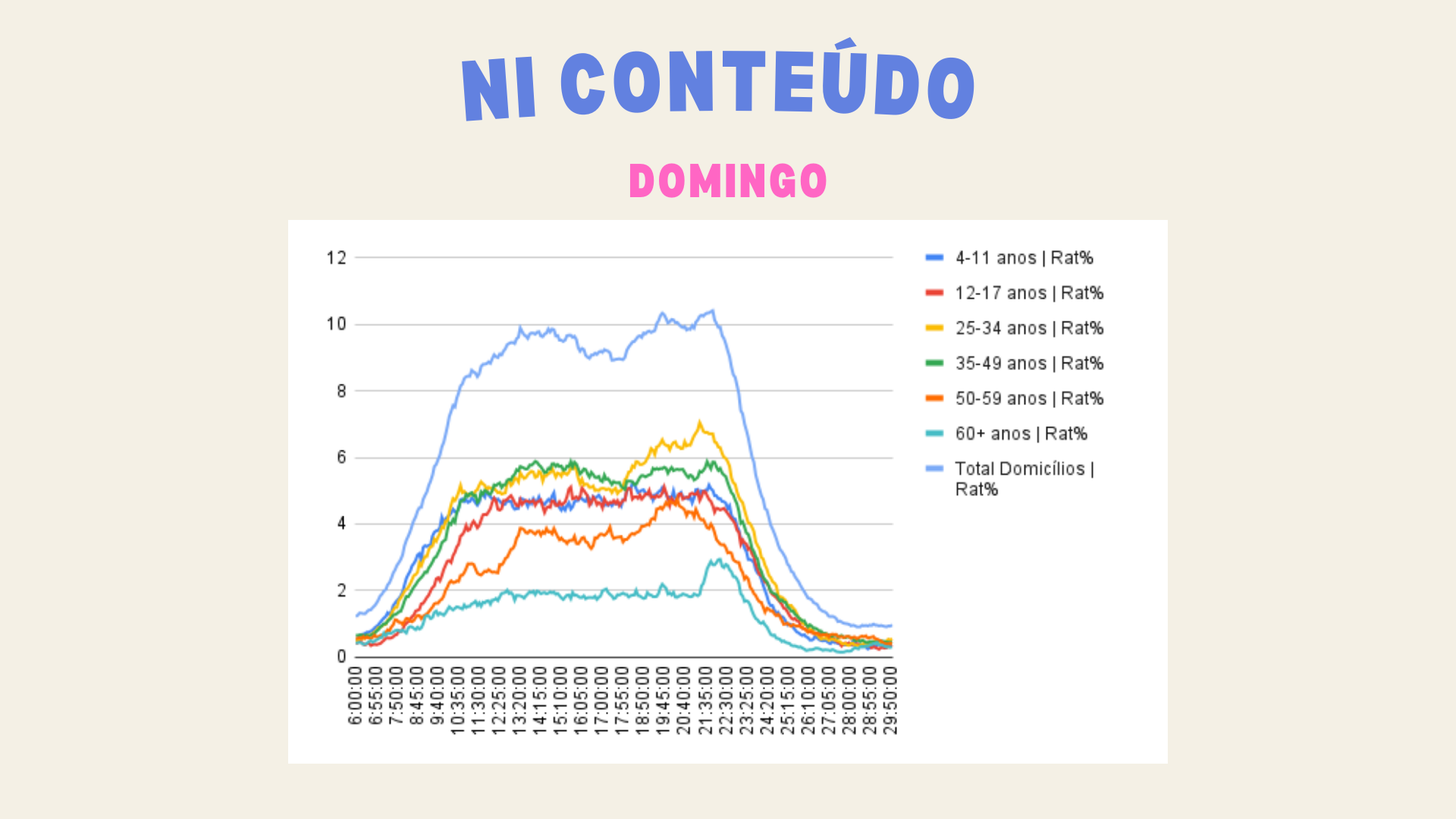
****

****

**Conteúdo Não Identificado:**

****

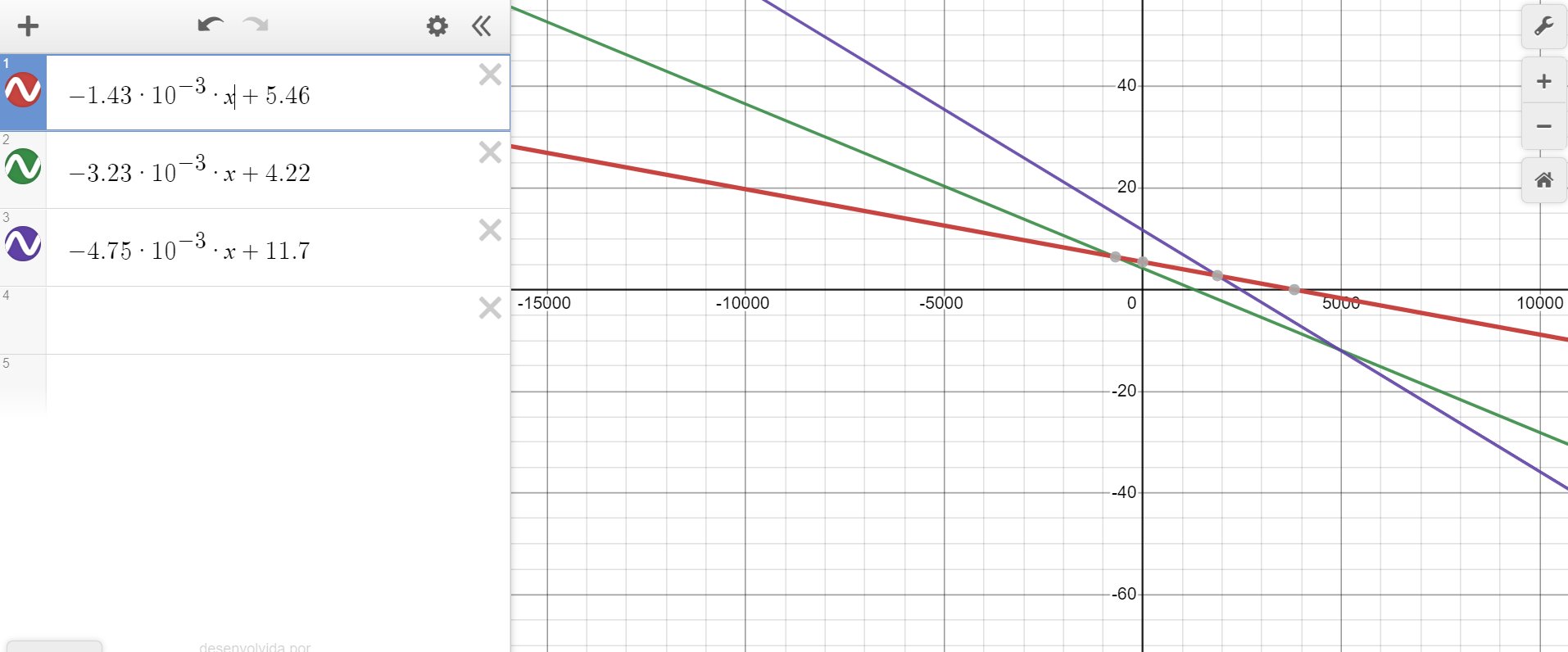
****

****

1. Descrição da predição desejada (“target”), identificando sua natureza (binária, contínua, etc.)

A natureza da predição desejada é contínua, uma vez que a saída dos dados será feita através da predição de um score de audiência.

Gráfico de tendência das 3 emissoras:



## 4.3. Preparação dos Dados

Descreva as etapas realizadas para definir os dados e os atributos descritivos dos dados (“features”) a serem utilizados. Essa descrição deve ser feita de modo a garantir uma futura reprodução do processo por outras pessoas, e deve conter:

a) Descrição de quaisquer manipulações necessárias nos registros e suas respectivas features.

b) Se aplicável, como deve ser feita a agregação de registros e/ou derivação de novos atributos.

c) Se aplicável, como devem ser removidos ou substituídos valores ausentes/em branco.

d) Identificação das features selecionadas, com descrição dos motivos de seleção.

Não deixe de usar tabelas e gráficos de visualização de dados para melhor ilustrar suas descrições.

IMPORTANTE: Crie tópicos utilizando a formatação “Heading 3” (ou menor) para que o Google Docs identifique e atualize o Sumário (é necessário apertar o botão Refresh no Sumário para ele coletar as atualizações)

## 4.4. Modelagem

Para a Sprint 3, você deve descrever aqui os experimentos realizados com os modelos (treinamentos e testes) até o momento. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

Para a Sprint 4, você deve realizar a descrição final dos experimentos realizados (treinamentos e testes), comparando modelos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

## 

## 4.5. Avaliação

Nesta seção, descreva a solução final de modelo preditivo, e justifique a escolha. Alinhe sua justificativa com a seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

## 4.6 Comparação de Modelos

# 5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo, e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

# 6. Referências

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

LUCK, Heloisa. **Liderança em gestão escolar**. 4. ed. Petrópolis: Vozes, 2010.

SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.

# Anexos

Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.