**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO**

GABRIEL FERNANDES RASQUINHO - SP3084094

JOSÉ MARCOS DE ASSIS - SP3086518

LUÍS GUSTAVO CRUZ - SP3082067

# PROJETO - ANÁLISE EXPLORATÓRIA DE DADOS

**SÃO PAULO**

**19/11/2023**

## SUMÁRIO

# 1. Introdução

Este trabalho concentra-se na realização de uma análise exploratória da base de dados originada do Sistema de Atenção à Saúde Indígena (SASI), especialmente do Módulo de Vigilância Alimentar e Nutricional (VAN). A base em questão oferece informações essenciais sobre a saúde das crianças indígenas, destacando-se como uma fonte valiosa para avaliar o estado nutricional desse grupo específico de pacientes.

No contexto da disciplina ESP1A5 - Estatística e Probabilidade, a relevância desse estudo reside na aplicação prática de conceitos estatísticos e probabilísticos para explorar minuciosamente os dados do VAN do SASI.

## 2. Objetivos

No escopo da disciplina ESP1A5, este estudo visa aplicar técnicas estatísticas e probabilísticas avançadas para uma exploração detalhada dos dados do Módulo de Vigilância Alimentar e Nutricional (VAN) do SASI. A ênfase principal recai na análise descritiva das variáveis, utilizando também conceitos estatísticos mais avançados para identificar padrões, tendências e possíveis correlações. A pesquisa propõe-se a não apenas aprofundar o entendimento de estatística e probabilidade, mas também aplicar esses conhecimentos na interpretação e compreensão de questões cruciais relacionadas à saúde indígena.

Este documento busca documentar de maneira abrangente todo o processo, desde a coleta inicial até o reconhecimento dos dados e a análise exploratória completa. Cada etapa será detalhada, proporcionando uma visão completa do tratamento dado aos dados desde sua origem até a interpretação final.

## 3. Dicionário de Dados

A seguir, apresenta-se o dicionário de dados que desempenha um papel crucial como guia para a análise exploratória empreendida. Esse conjunto de informações concede uma visão abrangente das variáveis contidas na base de dados, proporcionando, assim, uma compreensão sólida das possibilidades iniciais para a análise.

| Nome da Variável | Tipo | Descrição | Nome do Campo |
|---|---|---|---|
| DSEI_GESTAO | VARCHAR2 (100) | Descrição do Distrito Sanitário Especial Indígena | DSEI de Gestão |
| DS_POLO_BASE | VARCHAR2 (100) | Descrição do Polo Base | Polo Base |
| CO_MUNICIPIO_IBGE | VARCHAR2 (6) | Código IBGE do Município | Código do Município |
| NO_MUNICIPIO | VARCHAR2 (60) | Nome do Município sem acentos | Município |
| SG_UF | VARCHAR2 (2) | Sigla da Unidade da Federação | UF |
| CO_INDIO_DESIDENTIFICADO | VARCHAR | Código único que identifica o acompanhamento | Código do Individuo |

| | | | |
|---|---|---|---|
| TP_SEXO | VARCHAR2 (2) | Sexo do indivíduo: M – Masculino, F - Feminino | Sexo |
| DT_ATENDIMENTO | DATE | Data do Atendimento | Data do Atendimento |
| NU_PESO | NUMBER(6, 3) | Peso do indivíduo em quilos | Peso |
| NU_ALTURA | NUMBER(8, 3) | Altura do indivíduo em centímetros | Altura |
| IDADE_ATEND | NUMBER | Anos de vida do indivíduo no dia do atendimento | Idade em Anos |
| FAIXA_ETARIA | VARCHAR | Faixa etária calculada de acordo com a idade | Faixa Etária |
| TIPO_ALEITAMENTO | VARCHAR2 (60) | Tipo de aleitamento materno | Tipo de Aleitamento Materno |
| DS_PESO_IDADE | VARCHAR2 (50) | Descrição do peso por idade | Peso por Idade |
| DS_ESTATURA_IDADE | VARCHAR2 (50) | Descrição da estatura por idade | Estatura por Idade |
| DS_IMC_IDADE | VARCHAR2 (50) | Descrição do IMC por idade | IMC por Idade |

## 4. Preparação dos Dados

Na fase de preparação dos dados, realizou-se a coleta no portal do openDataSus. Após essa etapa, foi extraída uma amostra aleatória correspondente a 10% da base de dados original. Durante a análise da amostra, identificou-se uma coluna com alguns valores nulos, representando menos de 5% do total. Considerando a relevância desses dados para a análise, optou-se por substituir os campos N/A por "Sem Informação", garantindo que essa alteração não impactasse negativamente na qualidade da análise. Adicionalmente, efetuou-se a renomeação dos campos na amostra, visando facilitar tanto a análise quanto a interpretação dos dados.

# 5. Análise Exploratória

## 5.1. Visão Geral do Conjunto de Dados

A amostra do conjunto de dados sob análise engloba um total de **9.181** registros, abrangendo informações relativas a diversos indivíduos. Cada campo do conjunto segue uma formatação específica, alinhada às normas e critérios previamente estabelecidos. Essa abordagem visa assegurar a coerência e precisão das informações contidas em cada variável. A seguir, é apresentada uma amostra que ilustra a organização estruturada dos dados neste conjunto analisado.

DSEI : MARANHÃO
PoloBase : ZÉ DOCA
MunicipioIBGE : 210087
Municipio : ARAGUANA
UF : MA
CodIndio : 83CF7AA65DC0DC8B1B9821D47C84B0D487472932
Sexo : M
DtAtendimento : 18/5/2022 00:00:00
Peso : 13
Altura : 91
Idade : 2.00
FaixaEtaria : ENTRE 2 ANOS A 5 ANOS
TipoAleitamento : Alimentação Complementar
PesoIdade : PESO ADEQUADO PARA A IDADE
EstaturaIdade : ESTATURA ADEQUADA PARA A IDADE
IMCIdade : EUTROFIA

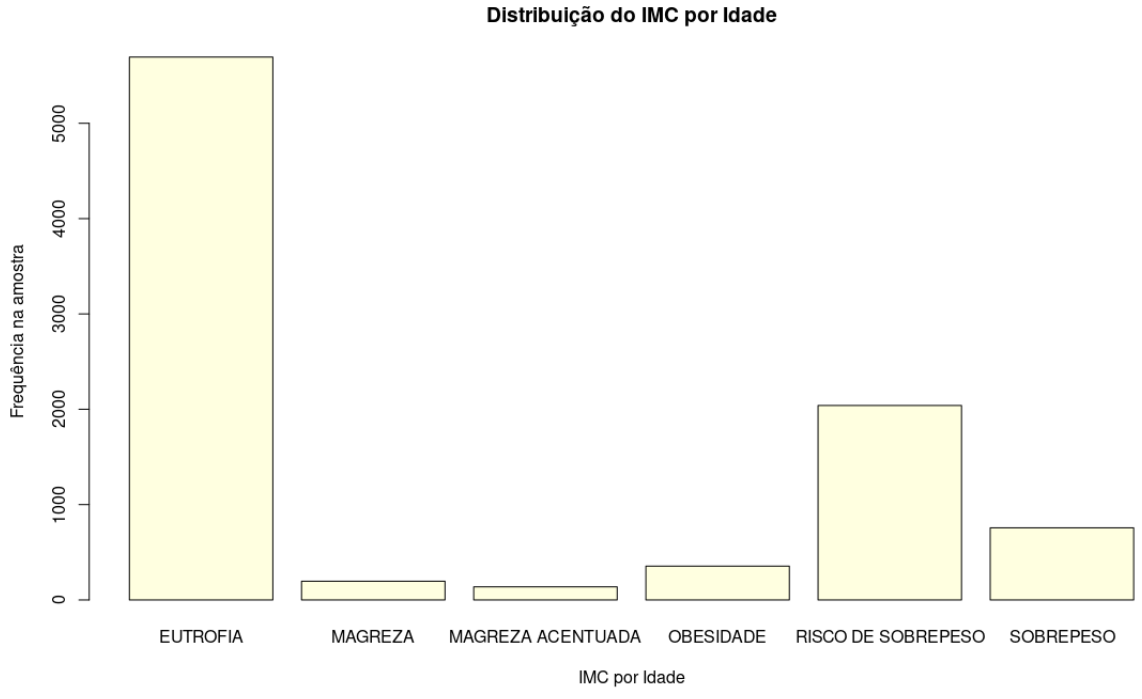
Classificação dos tipos de dados presentes na amostra analisada:
Tipo da variável DSEI : character
Tipo da variável PoloBase : character
Tipo da variável MunicipioIBGE : integer
Tipo da variável Municipio : character
Tipo da variável UF : character

Tipo da variável CodIndio : character
Tipo da variável Sexo : character
Tipo da variável DtAtendimento : character
Tipo da variável Peso : numeric
Tipo da variável Altura : numeric
Tipo da variável Idade : numeric
Tipo da variável FaixaEtaria : character
Tipo da variável TipoAleitamento : character
Tipo da variável PesoIdade : character
Tipo da variável EstaturaIdade : character
Tipo da variável IMCIdade : character

## 5.2. Distribuição de Frequências

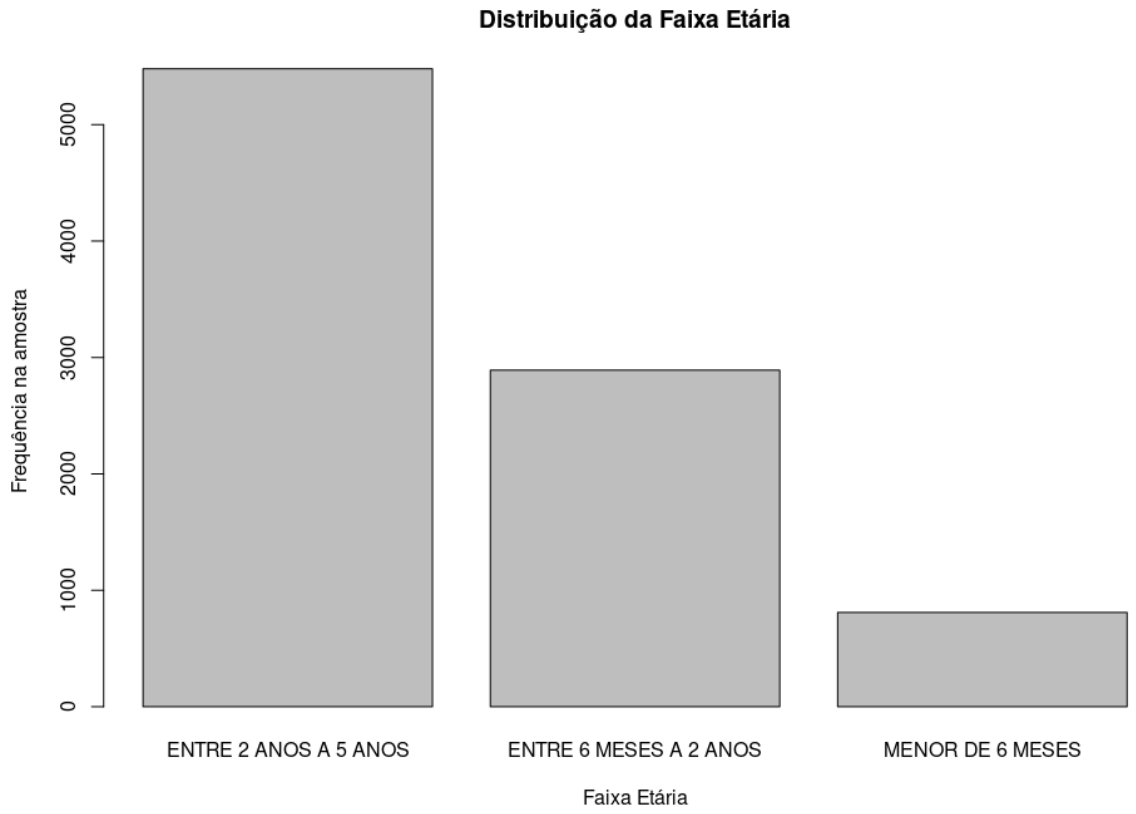
A presente análise exploratória tem como objetivo compreender o estado nutricional de crianças indígenas, destacando variáveis essenciais, tais como Faixa Etária, Tipo de Aleitamento e IMC por Idade. Na sequência, são apresentados os valores e as distribuições de frequência das variáveis mais pertinentes à análise proposta. Para isso, algumas perguntas foram feitas para a base.

1. **Qual é a distribuição total de IMC presente na amostra?**



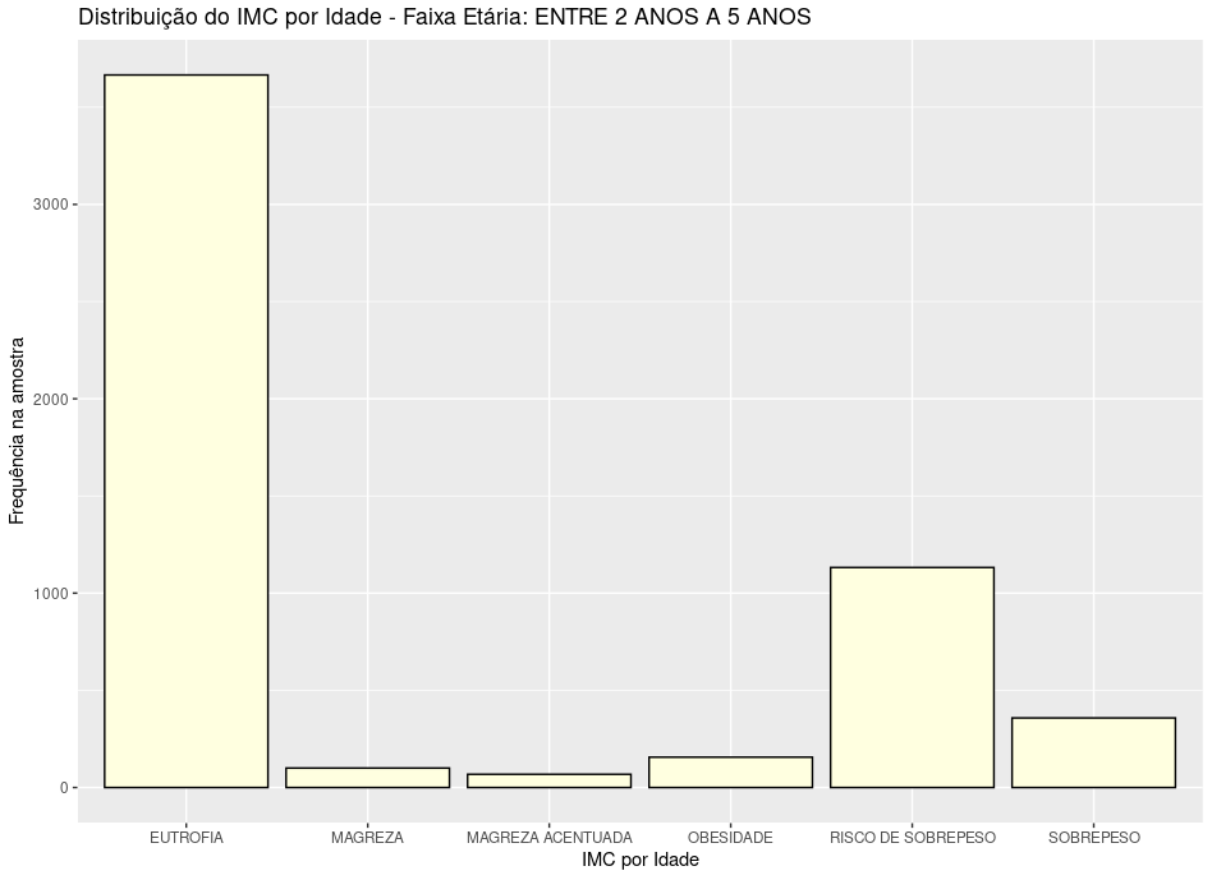
Analisando o histograma é possível notar que há uma grande frequência na variável EUTROFIA e que essa variável representa uma boa parte da amostra. Sendo uma frequência total de **5694**.

2. **Qual é a frequência de distribuição da faixa etária?**




Visualizando o histograma de distribuição da Faixa Etária, também é possível notar que boa parte dos dados da amostra são de indivíduos que possuem a faixa etária “ENTRE 2 ANOS A 5 ANOS” do total da amostra sendo no total **5480**.

3. **Qual é a distribuição do IMC na Faixa Etária de 2 anos a 5 anos de idade?**




Agora, analisando o histograma do IMC por faixa etária de 2 anos a 5 anos, é possível visualizar que boa parte dos indivíduos que estão entre 2 a 5 anos possuem o IMC para a Eutrofia com a frequência de **3666**.

4. **Com base nos dados coletados existe alguma relação entre as variáveis?**

Para verificar a relação entre as variáveis, foi realizado o teste de qui-quadrado. Para realizar o teste, primeiramente é tirada a média entre os IMCs por faixa etária de 2 anos a 5 anos pela frequência total da amostra.

**Saída de dados:**

Total da Amostra:  9181
Média Eutrofia: 0.3993029081799368
Média Risco de Sobrepeso: 0.1232981156736739
Média Sobrepeso: 0.038993573684783796
Média Magreza: 0.010892059688487093
Média Obesidade: 0.016991613114039864
Média Magreza Acentuada: 0.0074066005881712
Soma das Médias: 0.5968848709290927

Após tirar a média, temos as nossas proporções esperadas e realizamos o teste de qui-quadrado.

A saída do teste de qui-quadrado foi:
Teste de Qui-Quadrado: 238.15165939303466
Valor p: 1.932076128835257e-52

Com a saída do teste, é notável que o qui-quadrado obteve um valor alto, isso sugere que as variáveis estão associadas de alguma maneira.

O valor de p baixo indica que é extremamente improvável que a associação observada seja devida ao acaso.

Com base nos resultados, concluímos que há uma associação estatisticamente significativa entre as variáveis em sua tabela de contingência. A associação não é simplesmente devida ao acaso, há uma relação estatisticamente significativa entre as categorias das variáveis.

5. **Qual é a porcentagem dos IMCs presentes na amostra que estão na faixa etária de 2 a 5 anos de idade?**

Com teste de qui-quadrado realizado, prosseguimos para o cálculo de porcentagem sobre as médias que foram tiradas.

**Saída de dados:**

Porcentagem: IMC na faixa etária de 2 a 5 anos
PC Eutrofia: 39.93%
PC Risco de Sobrepeso: 12.33%
PC Sobrepeso: 3.90%
PC Magreza: 1.09%
PC Obesidade: 1.70%
PC Magreza Acentuada: 0.74%
Valor Total: 59.69%

Notasse que a faixa etária de 2 a 5 anos representa 59,69% da nossa amostra e que o IMC de Eutrofia na faixa etária de 2 a 5 anos representa quase 40% da amostra total.

6. **Qual é a frequência de indivíduos que possuem o peso adequado para a idade na faixa etária de 2 anos a 5 anos e qual é a frequência das estaturas em porcentagem?**

Prosseguindo com a análise, verificaremos as variáveis de altura e peso na faixa etária de 2 anos a 5 anos com indivíduos que possuem o peso adequado para a idade.

A contagem da variável PESO ADEQUADO PARA A IDADE mostrou uma frequência de **4982** em nossa amostra. A partir disso olharemos para as estaturas que possuem as seguintes variáveis: **ESTATURA ADEQUADA PARA A IDADE, BAIXA ESTATURA PARA A IDADE, MUITO BAIXA ESTATURA PARA A IDADE.**

**Observando a frequência em porcentagem das estaturas temos a seguinte saída de dados:**

Total: 9181
Estatura Adequada para a idade: 42.75%
Baixa Estatura para a idade: 8.71%
Muito Baixa Estatura para a idade: 2.80%
Analisando a saída, algumas conclusões podem ser tiradas. De 9181 dados que equivalem a 100% da nossa amostra, 42,75% apresentam possuir a estatura adequada para a idade e 11.51% não. Notasse que quase metade da amostra tende a possuir uma estatura adequada para a idade.

## 7. Quais são os tipos de variáveis presentes na coluna de altura?

Prosseguindo com a análise, verificamos que a coluna de Altura possui variáveis numéricas contínuas e é gerado uma visualização dos valores distintos dessa amostra.

**Saída de dados:**

```
[ 72.   73.   74.   75.   75.5  76.   76.5  77.   77.5  78.   79.   79.5
  80.   80.5  81.   81.5  82.   82.5  83.   83.5  84.   84.5  85.   85.5
  86.   86.5  87.   87.5  88.   88.5  89.   89.5  90.   90.5  91.   91.5
  92.   92.5  93.   93.5  94.   94.5  95.   95.5  96.   96.5  97.   97.5
  98.   98.5  99.   99.5 100.  100.5 101.  101.5 102.  102.5 103.  103.5
 104.  104.5 105.  105.5 106.  106.5 107.  107.5 108.  108.5 109.  110.
 110.5 111.  111.5 112.  113.  113.5 114.  115.  116.  117.  118.  119.
 120.  122.  123.  132.  134. ]
```

## 8. Com base nos dados da coluna altura, existe algum outlier?

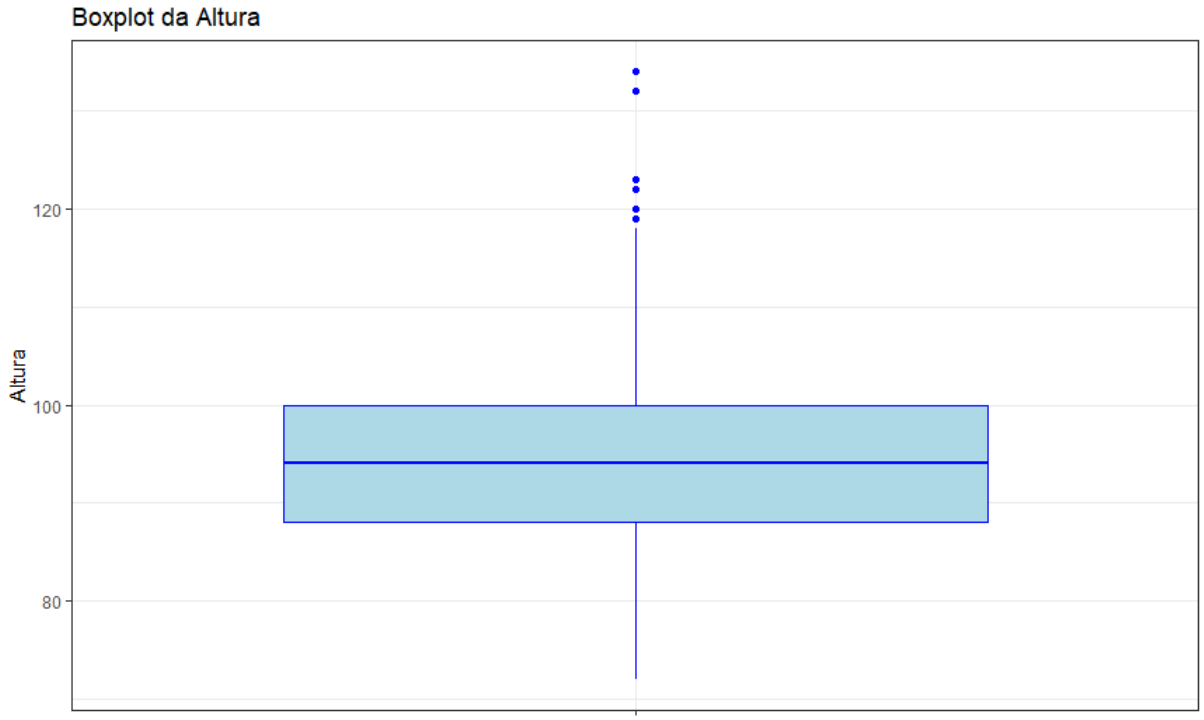
Após a verificação dos dados distintos presentes na coluna, analisaremos o intervalo entre quartis da Altura para verificar se existem outliers.

**Saída de dados:**

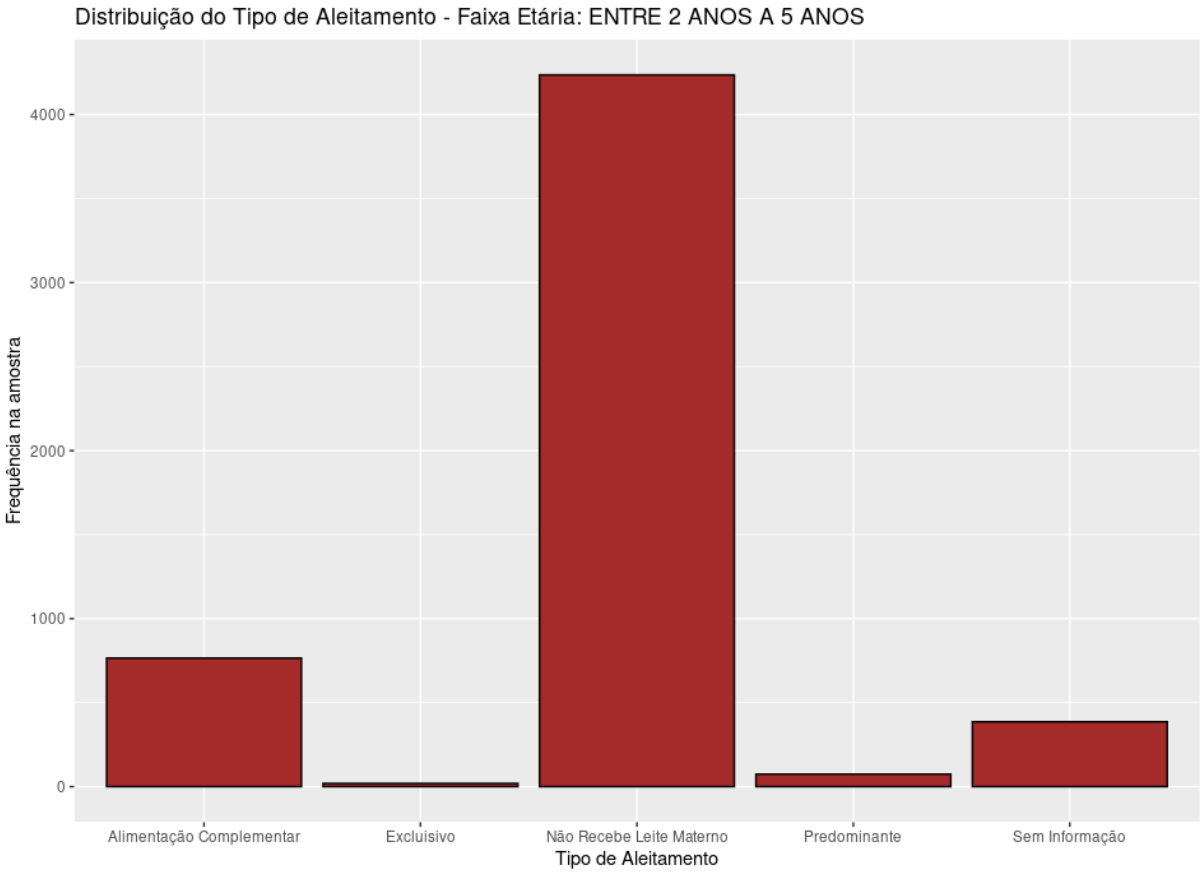
Mediana: 94.0
Q1: 88.0
Q2: 100.0
IQR: 12.0
Limite Inferior: 70.0
Limite Superior: 118.0

Outliers encontrados: 119.0, 120.0, 122.0, 123.0, 132.0, 134.0

Após encontrarmos os outliers, é gerado um boxplot para facilitar a visualização dos outliers presentes na amostra.




**9. Qual o tipo de aleitamento mais frequente para a faixa etária: ENTRE 2 ANOS A 5 ANOS ? Está de acordo com o esperado ?**



O histograma de barras acima realiza uma distribuição de dados em categorias específicas. Ele exibe cinco categorias de aleitamento ao longo do eixo horizontal(X), que são: Alimentação Complementar, Exclusivo, Não recebe Leite Materno, Predominante e Sem Informação. O eixo vertical (Y) por sua vez representa a frequência de cada categoria dentro da amostra.

**Saída de dados:**

Alimentação Complementar: 765
Exclusivo: 19
Não Recebe Leite Materno: 4236
Predominante: 74
Sem Informação: 185

**Com base nele, podemos realizar algumas análises:**

**Alimentação Complementar**: Apresenta baixa frequência, indicando que a alimentação complementar ao leite materno não é comum nesta faixa etária.

**Exclusivo**: O registro desta categoria é mínimo, apenas 19 crianças desta faixa etária se encaixam neste aleitamento, o que é esperado, pois a alimentação exclusiva por leite materno geralmente termina antes dos 2 anos de idade.

**Não Recebe Leite Materno**: A categoria mais frequente, refletindo a prática comum de desmame nesta faixa etária.

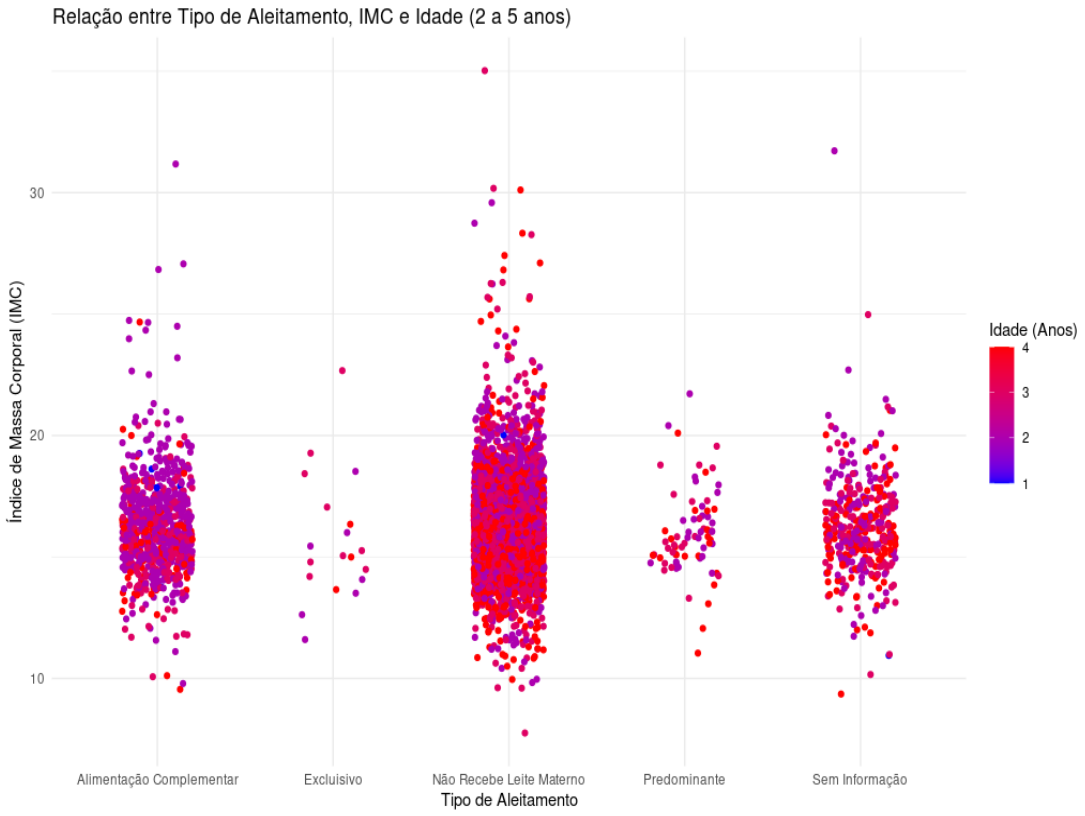
**Predominante**: Assim como a alimentação complementar, apresenta baixa frequência, sugerindo que poucas crianças ainda recebem leite materno como parte principal de sua alimentação.

**Sem Informação**: Há uma quantidade significativa de dados faltantes ou não especificados nesta categoria.

Dessa forma, conclui-se que a maioria das crianças de 2 a 5 anos na amostra analisada não recebe leite materno, o que está alinhado com as diretrizes de

alimentação infantil que recomendam o desmame e a introdução de uma dieta mais variada nessa faixa etária. A quantidade de dados faltantes na categoria “Sem Informação” ressalta a necessidade de coleta de dados mais rigorosa.

10. Existe algum tipo de relação entre o IMC das crianças com o tipo de aleitamento, na faixa de 2 a 5 anos?




O gráfico gerado é um jitter plot, que é uma versão do gráfico de dispersão que ajuda a visualizar melhor os dados quando há muitos pontos. No jitter plot, os pontos são ajustados no eixo x por uma pequena quantidade aleatória para minimizar a sobreposição, tornando mais fácil ver a distribuição dos dados.

No gráfico, o eixo X representa os diferentes tipos de aleitamento, e o eixo Y representa o Índice de Massa Corporal (IMC) das crianças. A cor de cada ponto é determinada pela idade da criança, com a escala de cores indo de azul (idade mais baixa) a vermelho (idade mais alta). Isto significa que, para cada tipo de aleitamento,

você pode ver uma dispersão de pontos que representam o IMC das crianças nessa categoria, com a cor indicando a idade da criança.

**Aqui estão alguns pontos a considerar ao interpretar o gráfico:**

**Distribuição do IMC:**

O IMC varia significativamente dentro de cada tipo de aleitamento, como indicado pela dispersão vertical dos pontos. Isso mostra que há uma variedade de valores de IMC entre as crianças, independentemente do tipo de aleitamento.

Alguns valores de IMC particularmente altos nas categorias “Alimentação complementar”, “Não recebe leite materno”, e “Sem informação” podem ser considerados outliers, já que se destacam acima do restante dos dados. Isso pode indicar casos de sobrepeso ou obesidade, ou resultado de erros de entrada de dados ou medições anormais que requerem verificação.

**Relação com a Idade:**

A distribuição de cores indica que há uma mistura de idades dentro de cada tipo de aleitamento, mas algumas tendências de cor podem ser observadas. Por exemplo, é possível notar uma alta frequência de crianças de 3 a 4 anos, com o IMC entre 10 e 20 que não recebem leite materno. Ao mesmo tempo em que cresce a presença de crianças de 2 anos na faixa de IMC 20, o que pode conter uma relação com este tipo de aleitamento nesta idade já que estão bem agrupados. Também é possível notar a forte presença de crianças de 2 anos na categoria de “Alimentação complementar”, novamente mantendo o IMC entre 10 a 20.

A intensidade e o agrupamento das cores sugere uma tendência na idade predominante associada a certos valores de IMC dentro de um tipo de aleitamento.
No caso atual, “Não recebe leite materno” tem um domínio maior a partir dos 3 anos. Já a “Alimentação complementar” reflete uma presença maior de crianças de 2 anos. As demais categorias, “Exclusivo”, “Predominante” e “Sem informação”, para esta faixa etária, apresentam uma variação maior com relação a idade.

**Análise por Categoria de Aleitamento:**

"Alimentação Complementar" e "Não Recebe Leite Materno" tem uma ampla gama de IMC, indicando uma grande diversidade na condição nutricional das crianças.

A categoria "Exclusivo" parece sub-representada, o que faz sentido, pois a amamentação exclusiva é geralmente recomendada apenas até os 6 meses de idade. "Predominante" e "Sem Informação" têm pontos dispersos, mas sem um padrão claro de distribuição de idade ou IMC.

## 6. Conclusão

A análise exploratória dos dados provenientes do Sistema de Atenção à Saúde Indígena (SASI) proporcionou valiosas percepções acerca do estado nutricional das crianças indígenas, com foco na faixa etária entre 2 e 5 anos. A decisão estratégica de concentrar a investigação nesse intervalo foi respaldada pela expressiva representatividade, abrangendo 59,69% dos registros disponíveis. Entre os resultados salientes, destaca-se a prevalência de eutrofia, atingindo aproximadamente 39,93% da população estudada. Essa condição sugere uma saúde física adequada, enfatizando a importância de compreender os padrões nutricionais específicos desse grupo etário.

A análise também evidenciou uma notável diversificação nos padrões de aleitamento, com especial destaque para a categoria "Não Recebe Leite Materno", alinhada ao processo natural de introdução de alimentos sólidos na dieta infantil. A variação significativa nos índices de Índice de Massa Corporal (IMC) entre os diferentes tipos de aleitamento aponta para a necessidade premente de investigações mais detalhadas, indicando potenciais casos de sobrepeso ou obesidade. Em resumo, os resultados obtidos representam um ponto de partida essencial para a compreensão da dinâmica nutricional específica das crianças indígenas, fornecendo direcionamentos valiosos para a implementação de ações e orientando futuras pesquisas na área de saúde infantil indígena.