Uma introdução gentil à Ciência de Dados

Kaike Wesley Reis, Júlia Bijos e Janaína Souza

2020-09-01

Contents


```
knitr::opts_chunk$set(error = TRUE)
```

Chapter 1

Capitulo 1

Chapter 2

Capitulo 2

Chapter 3

Visualização e Ciência de dados

O capítulo 2 apresenta a tabela como uma forma poderosa para estruturar e visualizar informações. No entanto, quando trabalhamos com enormes tabelas com uma imensa quantidade de linhas e colunas se torna difícil interpretar suas informações, não importa o quão organizadas elas estejam. Às vezes, é muito mais fácil interpretar essas informações através dos gráficos, conteúdo que será explorado no decorrer deste capítulo.

A construção e visualização gráfica é de extrema importância na área de ciência de dados, pois é a partir de um bom gráfico que podemos extrair ideias, hipóteses e um melhor entendimento a respeito de um tema ou uma pergunta. A importância desse tipo de análise pode ser expressa por um ditado popular bastante conhecido: "Uma imagem vale mais que mil palavras".

3.1 Objeto de estudo do capítulo

Para compreender a importância da análise gráfica e como utiliza-la corretamente, iremos buscar entender o perfil dos estudantes de Salvador que realizaram a prova do Exame Nacional do Ensino Médio (ENEM) no período de 2015 até 2019.

Porém, antes de qualquer coisa: O que é um **Perfil**? Esse termo é muito usado na estatística para **descrever determinado processo ou objeto de estudo, buscando entender características e padrões que o representa**. Para este caso em específico, vamos estudar os estudantes da cidade de Salvador utilizando os microdados do ENEM, publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), disponível ao público através deste link de acesso.

Como o termo **perfil** pode ser bem vasto e diversas características podem ser extraídas, é necessário concentrar essa análise em perguntas mais específicas para nortear o estudo. No decorrer deste capítulo, será explorado graficamente as seguintes questões:

- A quantidade de estudantes que realizam o ENEM aumentou de 2015 para 2019 na capital bahiana?
- Como é a distribuição de estudantes em Salvador por cor/raça? Conseguimos identificar algum padrão para esses valores?
- Na dita era da informação, onde tudo está conectado, como está o acesso dos estudantes a internet em suas residências? E a computadores pessoais?
- O tipo de escola (pública ou privada) pode influenciar nas notas dos estudantes neste exame?

A compreensão desses dados é de suma importância para compreender melhor o perfil dos estudantes de Salvador que possuem o ENEM como uma oportunidade de acesso, as vezes única, ao ensino superior no Brasil.

3.2 Gráfico de barra

O **Gráfico de barras** é uma forma bastante comum e versátil de visualização na área de ciência de dados. Ele pode ser utilizado tanto com variáveis categóricas quanto numéricas para expressar grandezas. A Figura abaixo apresenta uma de suas utilizações: demonstrar grandezas numéricas.

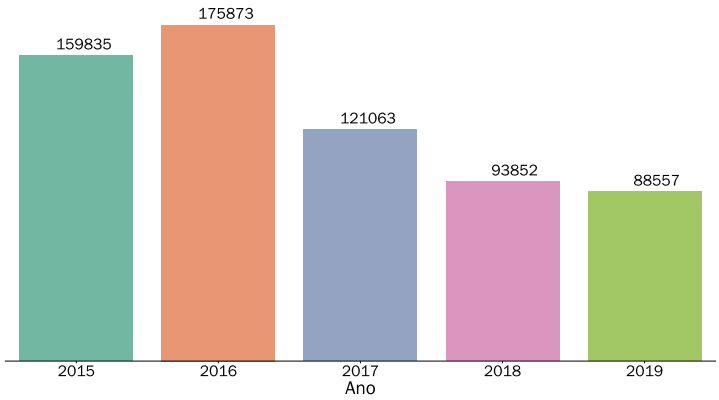

Figure 3.1: Quantidade de estudantes que se inscreveram o ENEM na capital bahiana

Na Figura 3.1 é apresentado a quantidade de estudantes que realizaram o ENEM de 2015 até 2019 na capital bahiana. É possível notar uma queda drástica na

participação de estudantes entre os períodos de 2016 até 2019. Apesar de simples e direto, a análise desse mesmo resultado através de uma tabela pode se mostrar confusa:

Ano	Número de estudantes em Salvador
2015	159835
2016	175873
2017	121063
2018	93852
2019	88557

Note que ao visualizar a Tabela, nenhuma informação visual é passada para destacar os anos com mais ou menos participantes. Além disso, ela contém as mesmas informações demonstradas na Figura 3.1, porém com uma diferença: através da visualização gráfica fica muito mais claro a queda de inscrições no ENEM de 2016 até 2019. O gráfico de barras apresenta uma característica muito importante relacionado ao tamanho das barras: elas crescem proporcionalmente de acordo as grandezas que elas se referem, ou seja, quanto maior o valor maior será sua barra. Comumente essas barras apresentam a mesma largura neste tipo de gráfico.

É através da Figura 3.1 que podemos responder a primeira pergunta:"**A quantidade de estudantes que realizam o ENEM aumentou de 2015 para 2019 na capital bahiana?**" E a resposta é não. Apesar do número de estudantes crescer de 2015 para 2016, ocorre uma queda drástica até 2019, chegando a diminuir pela metade o número de inscrições em Salvador.

Essa resposta pode te levar a um questionamento mais profundo como "O que realmente motivou essa queda?". Infelizmente encontrar a resposta para este questionamento não é trivial, requer pesquisas mais específicas a cerca do tema, o que foge do escopo deste capítulo. Todavia, é interessante refletir como a partir de um simples gráfico, podemos alcançar perguntas ainda mais complexas.

Agora que respondemos a primeira questão, podemos perceber que a pergunta "**Como é a distribuição de estudantes em Salvador por cor/raça? Conseguimos identificar algum padrão para esses valores?**" está bastante relacionada ao seu resultado. Inicialmente para entender essa relação, precisamos entender o que seria essa distribuição de raças no questionário no ENEM. Trata-se de uma pergunta que busca entender como o estudante se classifica em relação a sua cor. Essa pergunta possui 7 respostas possíveis:

- Não declarado
- Pardo
- Preta
- Branco

- Amarelo
- Indígena
- Opção de não apresentar tal informação

Como foi explicado no Capítulo 2, esse questionamento pode ser definido como uma variável categórica dado a quantidade finita de opções apresentadas. Esse questionamento está bastante relacionado com a primeira questão, pois a quantidade total de estudantes pode alterar essa distribuição, aumentando ou diminuindo a depender das categorias.

Como tivemos uma diferença tão grande entre o número de inscritos em 2016 e 2019 demonstrado na Figura 3.1, uma análise mais aprofundada nesses dois anos podem trazer resultados interessantes para responder nosso segundo questionamento:

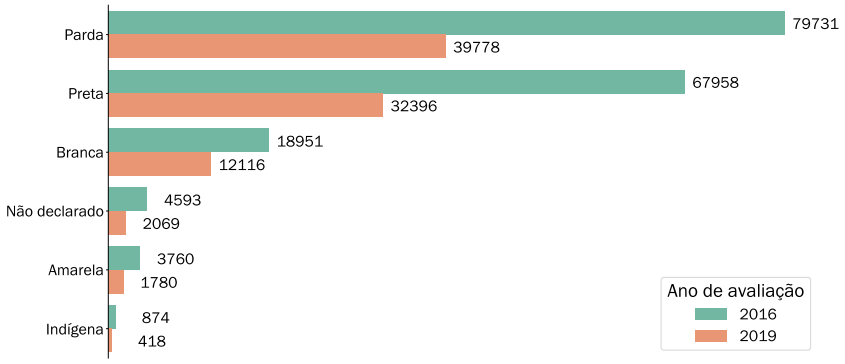

Figure 3.2: Distinção de estudantes inscritos por cor/raça da cidade de Salvador para os anos de 2016 e 2019

Através da Figura 3.2, é apresentado os valores absolutos da quantidade de estudantes que realizaram o ENEM em cada ano identificados pela sua raça. Note que a grande queda encontrada na Figura 3.1 se reflete neste gráfico também: Em comparação a 2016, todas as categorias apresentaram valores menores. Por exemplo, a quantidade pessoas pardas que realizaram o ENEM caiu quase pela metade, assim como as pessoas auto-declaradas como preta. Além disso, podemos notar uma baixíssima quantidade de pessoas indígenas/amarela que realizam este exame e que em sua grande maioria, os estudantes da capital bahiana se declaram como pardos e negros.

Essa situação já era esperada e reflete uma realidade já conhecida: Segundo o Instituto Brasileiro de Estatística e Geografia (IBGE), em uma pesquisa realizada

em 2017, Salvador é considerada a capital mais preta do brasil, onde 8 em cada 10 moradores se autodeclaravam de cor preta ou parda.

Note que a Figura 3.2 demonstra também a principal função do gráfico de barras: dimensionar variáveis categóricas de acordo a frequência de suas categorias. **Frequência** para uma variável categórica pode ser definido como a quantidade de vezes que ela é representada, podendo ser dividida em dois tipos: absoluta e relativa.

A frequência absoluta se trata da representação da quantidade de vezes que cada categoria ocorre. Este tipo de frequência é trabalhada na Figura 3.2, onde apresentamos a quantidade de estudantes por cor/raça que realizaram o ENEM nos anos de 2016 e 2019. Ainda na Figura 3.2, conseguimos notar que todas as categorias apresentaram uma queda na quantidade de estudantes que realizaram em 2016 para 2019, mas e se quisermos comparar este valores ainda utilizando um gráfico de barras, seria possível?

Uma boa forma para comparar essas frequências absolutas distintas seria através do segundo tipo de frequência apresentada anteriormente: a frequência relativa. A frequência relativa é definida como uma proporção entre o valor que você quer estimar e o valor máximo esperado. Podemos formular este conceito da seguinte forma:

$$Frequência\ Relativa(\%) = 100 * \binom{Valor\ para\ comparar}{Valor\ máximo}$$

Note que não foi mencionado o valor 100 presente na fórmula. Ele é apresentado para tornar o resultado da frequência relativa em porcentagem. Para compreender melhor este conceito apresentado, vamos continuar respondendo a segunda questão utilizando agora este novo aprendizado:

A Figura 3.3 pode ser vista como uma extensão da Figura 3.2, utilizando a frequência relativa para apresentar uma informação implícita: a proporção dos estudantes que fizeram o ENEM em 2019 em comparação a quantidade que realizaram o exame em 2016. Transcrevendo a fórmula da frequência relativa apresentada anteriormente, temos:

$$Frequência\ Relativa(\%) = 100 * \binom{Quantidade\ de\ estudantes\ \ realizaram\ o\ ENEM\ em\ 2019}{Quantidade\ de\ estudantes\ \ realizaram\ o\ ENEM\ em\ 2016}$$

Como nos é apresentado uma proporção, podemos ler o gráfico de barras apresentado na Figura 3.3 como sendo **a quantidade de estudantes que fizeram a prova em 2019 em relação a quantidade que realizou a prova em 2016**.

Podemos identificar, por exemplo, que com exceção dos estudantes autodeclarados de cor branca todas as outras raças apresentaram uma proporção de aproximadamente 50%, ou seja, o número de estudantes pardos, pretos, amarelos, indígenas e não declarados caíram pela metade em comparação ao

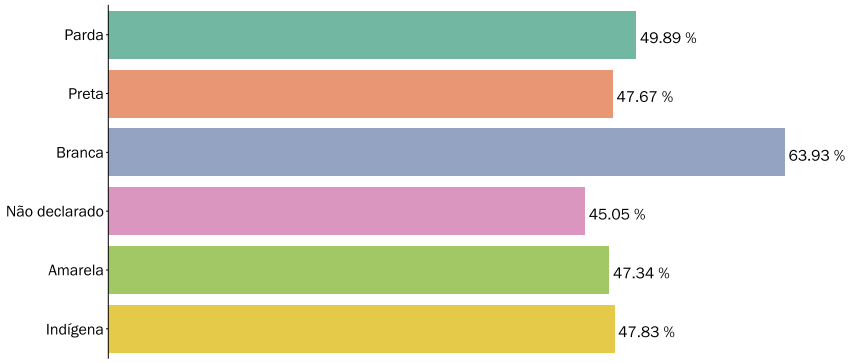

Figure 3.3: Comparação entre os estudantes inscritos de Salvador por cor/raça para 2016 e 2019

ano de 2016. Esta informação intensifica ainda mais o resultado apresentado na Figura 3.1.

Através da análise do gráfico de barras conseguimos avaliar dois questionamentos de uma só vez! Porém para analisar como esses resultados ocorreram de 2016 até 2019 ao invés de dois anos separados, qual seria o melhor tipo de gráfico? Iremos explora-lo na próxima seção deste capítulo.

3.3 Gráfico de tendências

Para responder com mais detalhes os dois questionamentos iniciais trazidos na seção anterior:

- A quantidade de estudantes que realizam o ENEM aumentou de 2015 para 2019 na capital bahiana?
- Como é a distribuição de estudantes em Salvador por cor/raça? Conseguimos identificar algum padrão para esses valores?

Vamos usar o **gráfico de tendências**. Este tipo de gráfico trata a visualização de uma coleção de observações realizadas ao longo do tempo para acompanhar um evento ou processo.Por se tratar de uma coleta sequencial, ou seja, feita uma após a outra torna o fator de ordenação fundamental: importa saber se determinada observação ocorreu antes ou depois de determinado evento.

Este conceito será importante para expandir a análise realizada apenas com os anos de 2016 e 2019 atráves do gráfico de barras, verificando como foi o comportamento do número de estudantes e sua distribuição por raça nesse

período como um todo, sem precisar olhar para apenas dois anos separadamente.

Porém, antes de mergulhar na análise gráfica desses dois questionamentos é importante explorar mais um conceito novo: o plano cartesiano.

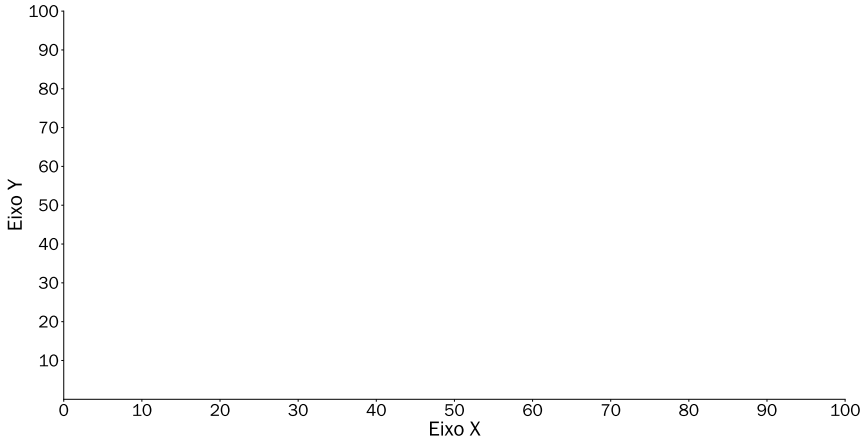

Figure 3.4: Plano cartesiano simplificado

A Figura 3.4 apresenta um plano cartesiano simplificado. São definidos dois eixos principais: o eixo X e o eixo Y. Cada eixo pode demonstrar o comportamento de uma variável desejada: Para o eixo X, ao aumentarmos o valor se move para a direita e ao diminuir o valor se move para a esquerda, já para o eixo Y, ao aumentarmos o valor se move para cima e ao diminuir para baixo. Ter esse conceito em mente será importante para as análises futuras.

O gráfico apresentado na Figura 3.5 fortalece ainda mais a resposta trazida para o primeiro questionamento: o número de estudantes que realizaram este exame não vem aumentando nos últimos cinco anos. É mostrado uma queda acentuada de 2016 para 2019. Porém através da análise dessa tendência, vemos que a maior queda ocorre de 2016 para 2017 com uma diminuição de mais de 50 mil inscrições. Esse gráfico mostra que a tendência de queda no ENEM não ocorreu de forma abrupta de 2016 para 2019, mas de forma gradual já que a partir de 2016, os valores apenas diminuiram com 2019 sendo o menor deles.

Ainda neste gráfico, podemos extrair um conceito bem interessante referente a esta modalidade de visualização: o pico. O pico é definido como um aumento expressivo identificado em um determinado período em comparação aos demais. No nosso caso, o pico de inscrições no ENEM em Salvador ocorreu em 2016, pois é o maior valor verificado dentro deste intervalo de cinco anos.

A Figura 3.6 é endereçado ao segundo questionamento. Podemos notar que a tendência das duas primeiras curvas, referente as cores parda e preta dos

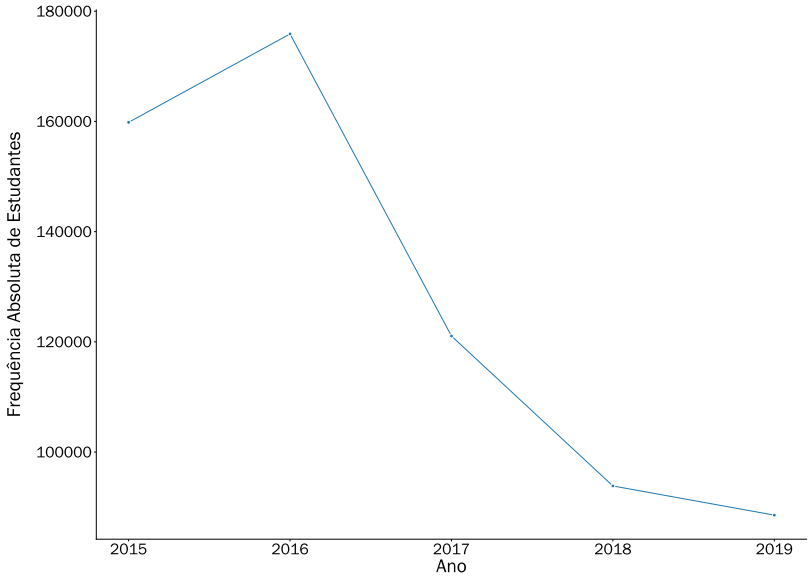

Figure 3.5: Quantidade de estudantes inscritos no ENEM na capital bahiana

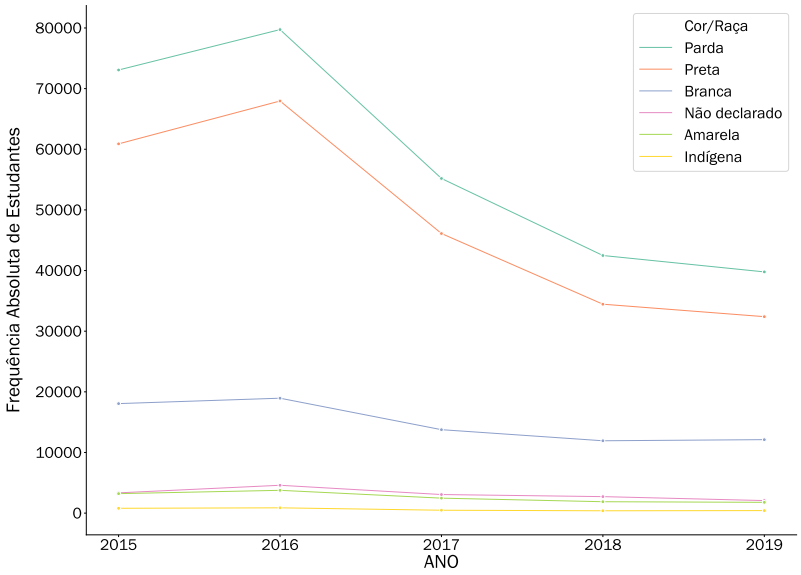

Figure 3.6: Tendência da quantidade de estudantes inscritos no ENEM por cor de 2015 até 2019

estudantes de Salvador, seguem um padrão similar ao que foi apresentado na Figura 3.5: Ocorre um pico em 2016 e a partir desse ano os valores decaem gradualmente. Porém este padrão fica bem claro para essas duas primeiras curvas, enquanto as outras se mostram aparentemente retilíneas, ou seja, não demonstram grande mudanças. Essa situação requer cuidados, pois podemos acreditar que para as outras opções não ocorreram nenhuma mudança. Essa divergência está relacionado a grandeza de cada curva: Valores maiores acabam esticando o gráfico, tornando valores pequenos menos representativos.

Para visualizar melhor e trazer uma melhor discussão a respeito do segundo questiomento, cada curva foi separada de acordo a raça que ela representa:

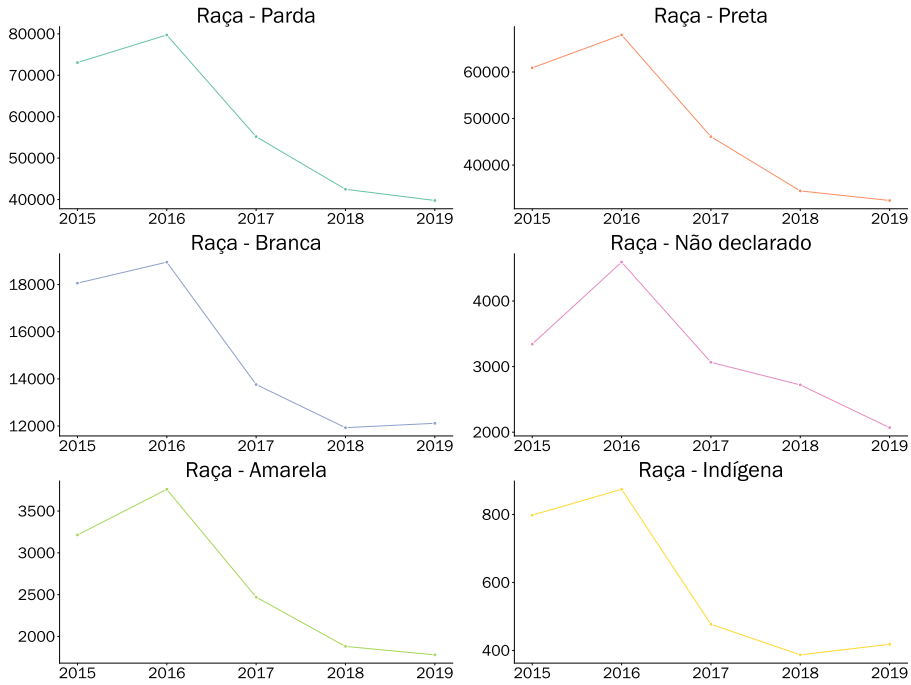

Figure 3.7: Tendência da quantidade de estudantes inscritos no ENEM particionado por cor de 2015 até 2019

Na Figura 3.7 conseguimos notar a diferença de grandezas que foi mencionado anteriormente ao visualizar o eixo Y (frequência absoluta de estudantes): para a cor parda, por exemplo, é possível enxergar valores próximos de oitenta mil estudantes enquanto para a cor amarela os valores ficam próximos de três mil e quinhentos estudantes mostrando assim uma grande disparidade. Ao separar os gráficos, cada um consegue ter sua própria escala, diferente da Figura 3.6 onde todos compartilhavam o mesmo eixo.

Respondendo ao segundo questionamento, verifica-se que as raças parda, preta,

amarela e as pessoas não declaradas seguem o padrão verificado dos estudantes inscritos em Salvador exposto na Figura 3.5: ocorre um pico em 2016, e a partir desse período os numéros de inscrições apenas caem. Todavía, para as pessoas de cor branca e indígena o padrão se mantém, porém em 2019 ocorre um leve aumento em comparação ao ano anterior. Esse aumento se difere apenas por ordem de grandeza entre si: enquanto para cor branca esse valor aumenta em torno de doze mil estudantes, para os indígenas eles aumentam em torno de 400 estudantes.

Através dessa análise gráfica conseguimos compreender e acompanhar como o número de inscritos no ENEM em Salvador veio se alterando nos últimos anos. Essa análise poderia ser utilizada para justificar tomadas de decisão na área da educação, buscando avaliar formas de aumentar a aderência dos estudantes para se inscrever no ENEM, através de programas sociais de fomento a educação, dado a importância deste exame para conseguir ingressar nas faculdades ou universidades da cidade ou país.

3.4 Gráfico de setores

Nas seções anteriores, conseguimos entender melhor o panorama dos estudantes de Salvador inscritos no ENEM nos últimos anos e como seus valores foram sendo alterados de acordo a quantidade e raça. Agora iremos avaliar o terceiro questionamento proposto no estudo de perfil: "**Na dita era da informação, onde tudo está conectado, como está o acesso dos estudantes a internet em suas residências? E a computadores pessoais?**". Essa pergunta é importante, pois acredita-se que hoje tudo está conectado e que o acesso a essas ferramentas, facilitadoras do aprendizado, é algo comum a todos, mas ... será? É possível que todos os estudantes do ENEM possuam fácil acesso as essas ferramentas no dias atuais? Iremos buscar responder este questionamento no decorrer deste capítulo.

Para isso será apresentado uma nova modalidade gráfica: o **gráfico de setores**. Este gráfico, usado comumente com variáveis categóricas, apresenta sua forma mais comum equivalente ao desenho de uma "pizza", onde cada fatia é referente a uma determinada categoria e seu tamanho é proporcional a sua representatividade. Para responder o primeiro questionamento, relacionado ao acesso da internet, vamos verificar um cenário mais atual e um cenário mais antigo, sendo respectivamente 2019 e 2015. Será que ocorreu melhorias no acesso à internet pelos estudantes do ENEM em Salvador?

Na Figura 3.8 é apresentado a frequência relativa dos estudantes com e sem acesso a internet de acordo ao total de estudantes soteropolitanos inscritos naquele ano. O uso da frequência relativa neste caso permite uma melhor comparação entre os anos e os resultados foram positivos: Em 2015 tinhamos 72,6% estudantes com acesso a internet e esse valor aumentou para 84,7% em 2019, mostrando uma melhora de 12,1%! Essa melhora é mostrada visualmente atráves do tamanho

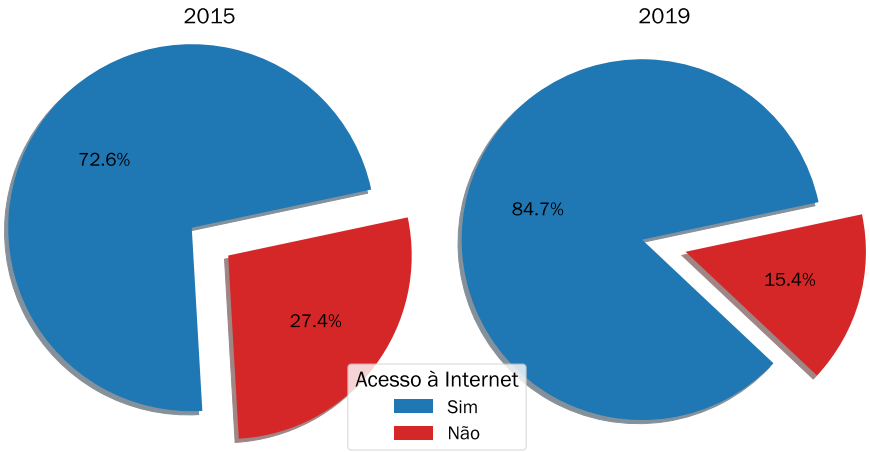

Figure 3.8: Porcentagem de estudantes inscritos no ENEM em Salvador com acesso a internet em 2015 e 2019

da fatia referente a resposta "Sim" de 2015 para 2019. Este resultado pode estar atrelado a diversos fatores como mais acessibilidade a esta ferramenta como a redução de custos, aperfeiçoamento dos projetos sociais de inclusão digital e etc. Deixamos a cargo do leitor buscar compreender os motivos que levaram a melhora nestes resultados.

Note que neste tipo de gráfico, ao utilizar frequência relativa, é necessário que a soma dos valores em todos os setores seja igual a 100%, isso não ocorre para o ano de 2015 devido a aproximação decimal utilizada de uma casa decimal.

Conseguimos encontrar parte da resposta do terceiro questionamento: "**Na dita era da informação, onde tudo está conectado, como está o acesso dos estudantes a internet em suas residências?**" E a resposta é que o acesso dos estudantes inscritos no ENEM à internet melhorou de 2015 para 2019, mas e o acesso a computadores pessoais em suas residências? Vamos utilizar novamente os anos de 2015 e 2019 para continuar esta pergunta:

Na Figura 3.9 podemos verificar que o questionário do ENEM em relação a esta pergunta possui 5 respostas representativas. Todavia, diferente do acesso a internet conseguimos avaliar que a fatia referente aos estudantes que possuem pelo menos um computador pessoal diminuiu de 61,5% em 2015 para 46,1% em 2019 enquanto o número de estudantes que não possuiam nenhum computador pessoal em sua residência aumentou de 28,1% em 2015 para 43,6% em 2019. A diferença entre essas duas proporções são semelhantes: enquanto uma fatia caiu 15,4% a outra aumentou 15,5% respectivamente. Esse resultado, atrelado ao encontrado na Figura 3.8 pode indicar que o acesso a internet realizado pelos estudantes podem surgir de outra fonte: celular. Essa narrativa é fortalecida com o seguinte argumento: De 2015 para 2019 ocorreu um aumento no acesso à

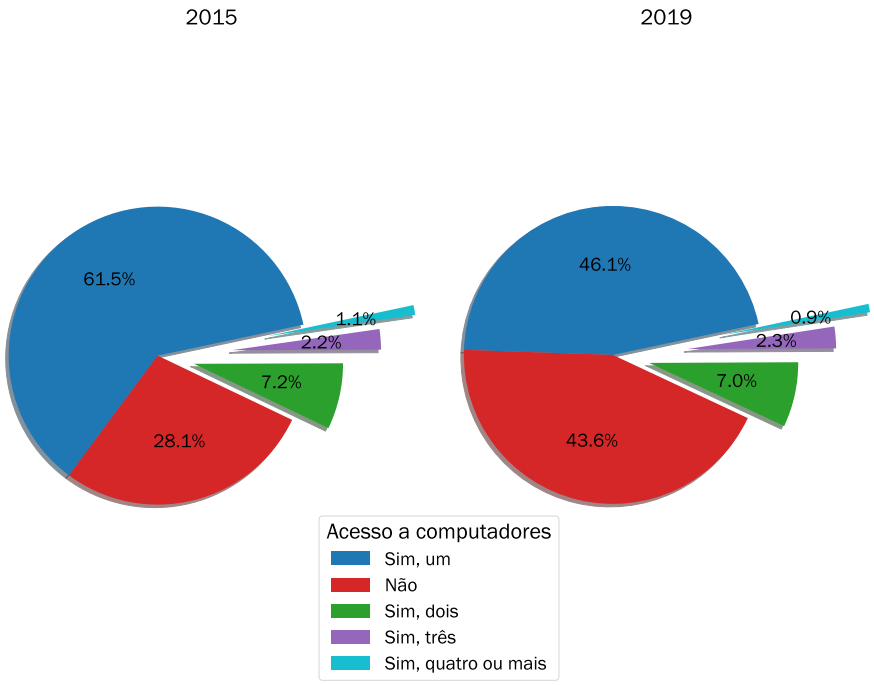

Figure 3.9: Porcentagem de estudantes inscritos no ENEM em Salvador com acesso a computadores pessoais em 2015 e 2019

internet, mas uma queda no acesso a computadores pessoais em domicílio.

Ainda na Figura 3.9, podemos avaliar que algumas fatias, referentes a estudantes com mais de um computador pessoal, são menos representativas dado o seu tamanho. Essa situação indica um dos problemas ao utilizar este tipo de visualização: quando uma variável possui muitas categorias ou categorias com pouca representatividade pode dificultar a visualização das informações para o leitor. Em casos como esse uma das recomendações é a utilização dos gráficos de barras. Porém existem outras formas de melhorar essa visualização: Como vimos que as categorias mais dominantes se referem aos estudantes sem ou com pelo menos um computador em casa, vamos juntar as categorias: "Sim, dois", "Sim, três" e "Sim, quatro ou mais" em uma só categoria: "Sim, mais de um". Será que isso pode melhorar a visualização do gráfico anterior?

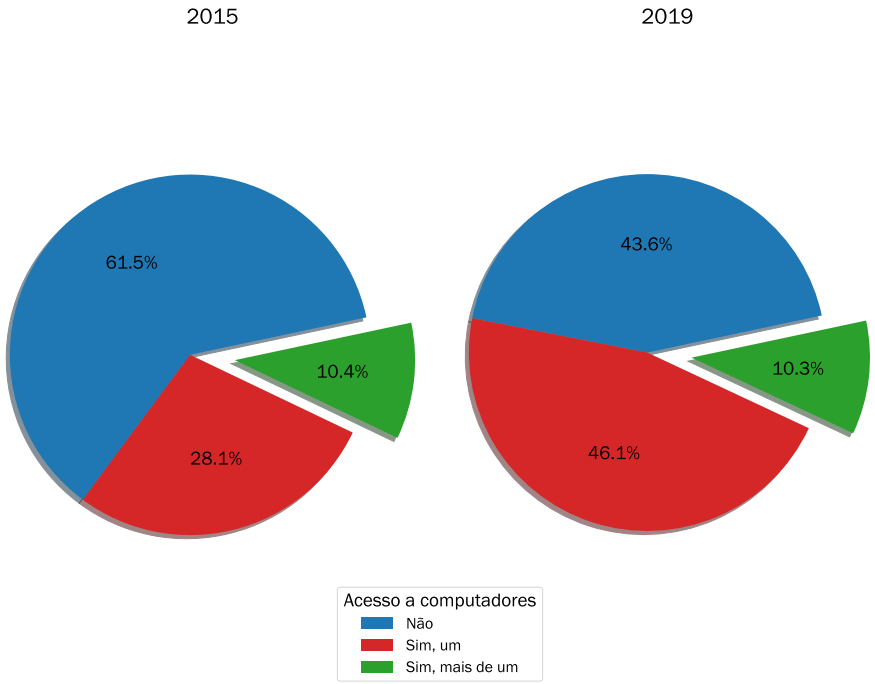

Figure 3.10: Porcentagem de estudantes inscritos no ENEM em Salvador com acesso a computadores pessoais em 2015 e 2019

Na Figura 3.10 é apresentado o resultado desta alteração. A confecção dessa nova categoria permitiu encontrar uma informação implícita no gráfico anterior: a proporção de estudantes com mais de um computador pessoal em casa se manteve praticamente constante de 2015 para 2019. Isso fortalece ainda mais a narrativa de uma queda na proporção de pessoas com pelo menos um computador pessoal em casa para a proporção de pessoas sem computador pessoal. Esse tipo de informação pode ser utilizada em programas sociais ou intervenções para

reverter este quadro.

Neste momento o leitor pode estar se questionando: Seria possível unir os dois resultados avaliados para este questionamento, acesso à internet e computador pessoal, em um só gráfico? Abaixo é mostrado que sim, podemos.

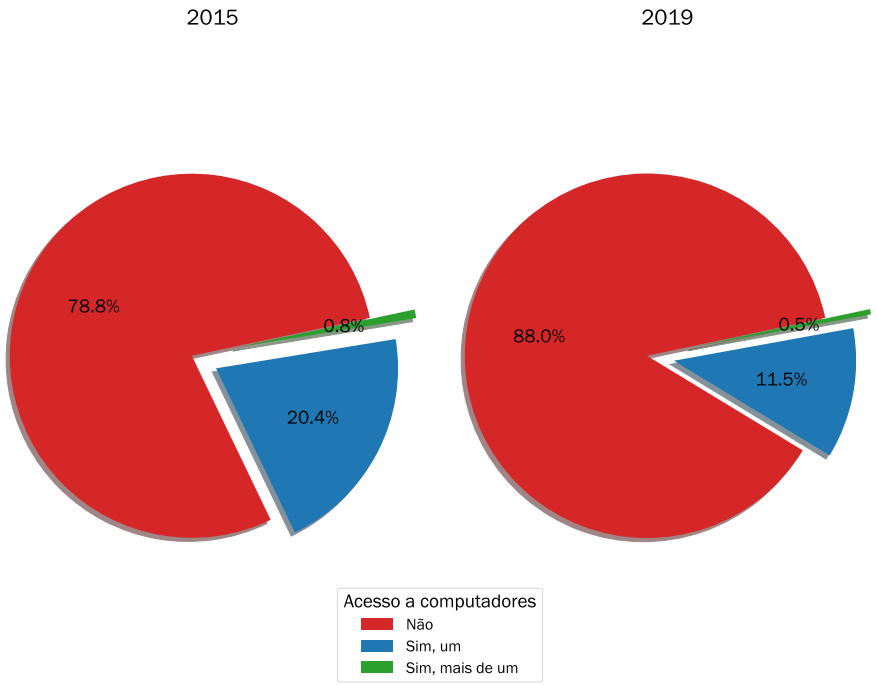

Figure 3.11: Porcentagem de estudantes inscritos no ENEM em Salvador sem acesso à internet em relação ao acesso a computadores pessoais em 2015 e 2019

Na Figura 3.11 são mostrada as proporções de estudantes de Salvador que não possui acesso à internet em 2015 e 2019 em relação ao acesso de computador pessoal. Podemos extrair deste gráfico algumas informações:

- Pode existir alguma incongruência na construção dessa base de dados, pois existem estudantes com mais de um computador pessoal, porém sem acesso à internet o que pode gerar questionamentos. Essa situação pode apresentar diversos motivos e uma das hipóteses mais plausíveis seria algum erro do estudante ao responder este questionário.
- É possível verificar que a maioria dos estudantes sem acesso à internet também não possui computadores pessoais em casa. Esta proporção cresce de 78,8% em 2015 para 88,8% em 2019 seguido pela queda da proporção de estudantes que possui pelo menos um computador pessoal em casa.

Essas informações podem indicar uma possível **correlação**, conceito que será estudado em capítulos futuros e de grande importância na área de ciência de

dados.

Assim é possível concluir o terceiro questionamento, que nessa era digital as situações melhoraram em partes: ocorreu um aumento, em termos proporcionais, de estudantes com acesso à internet, porém em contrapartida ocorreu um aumento de estudantes sem acesso a pelo menos um computador pessoal em suas residências o que pode dificultar sua navegação.

3.5 Gráfico de Dispersão

Até o momento conseguimos resolver graficamente 3 dos 4 questionamentos referente ao perfil dos estudantes de Salvador que realizaram o ENEM. Para responder o quarto questionamento: "**O tipo de escola (pública ou privada) pode influenciar nas notas dos estudantes neste exame?**" vamos utilizar uma nova ferramenta visual: o gráfico de dispersão. Para entender os motivos para escolhermos essa ferramentas precisamos antes apresentar seu conceito.

Gráficos de dispersão se tratam de representações usando duas ou mais variáveis através das coordenadas cartesianas para exibir valores de um conjunto de dados. Para ficar mais claro este conceito, vamos focar em responder o quarto questionamento utilizando as notas dos estudantes de Salvador no ano de 2019, considerando apenas aqueles que:

- Apresentaram uma pontuação maior que zero em todas as provas, com exceção no exame de Redação
- Definiram o tipo de colégio no ensino médio: público ou privado

Essas condições foram colocadas para evitar valores aberrantes nas análises. Além disso é importante mencionar que no ano de 2019, cerca de 75% dos estudantes de Salvador não responderam a questão referente ao tipo de colégio, logo as análises apresentadas aqui representam cerca de 25% dos estudantes inscritos no ENEM 2019.

Inicialmente, será mostrado um gráfico de dispersão para as provas da área de exatas: matemática e ciências naturais, mas não se assuste! O gráfico será explicado passo a passo.

Na Figura 3.12 é apresentado a nota dos estudantes de Salvador em matemática no eixo Y e no eixo X as notas em ciências naturais, destacando em cores o tipo de colégio: azul escola pública e em amarelo escola privada. Neste gráfico de dispersão são contemplados todos os estudantes que atenderam todos os requisitos expressos anteriormente, onde cada estudante é representado por um ponto de coordenada $(x,\ y)$ ou se preferir $(\textit{nota em ciências naturais},\ \textit{nota em matemática})$. Como o ENEM funciona por pontuação, o aluno que apresentar as maiores pontuações em todas as provas possui maior vantagem na escolha de um curso superior, ou seja, os estudantes com melhor rendimento são aqueles que se aproximam do canto superior direito. Apesar desta modalidade gráfica ser bem

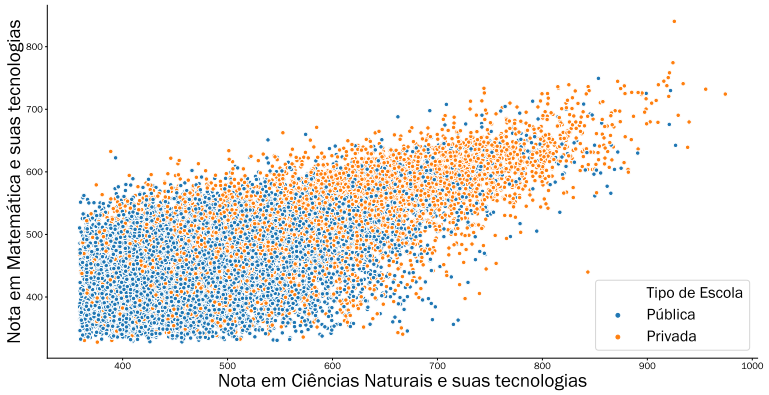

Figure 3.12: Relação entre nota de Ciências Naturais e Matemática particionado pelo tipo de escola em 2019

simples, ela pode trazer resultados interessantes e intuitivos.

Através da Figura 3.12 podemos verificar que a maioria dos estudantes de escolas públicas se localizam no canto inferior esquerdo, ou seja, estudantes com notas menores em ambas as provas e a medida que crescemos em ambos os eixos, mais dominante se tornam os estudantes de escolas privadas, mostrando um maior rendimento.

Além desta análise, no geral é possível verificar uma **tendência** crescente, onde ao aumentarmos a nota de matemática vemos que a maioria dos estudantes também aumentam a nota em ciências naturais. Compreender tendências deste tipo faz parte do dia a dia do cientista de dados, pois elas essas tendências são as mais comuns e intuitivas na natureza.

Na Figura 3.13 é apresentado dois padrões: em vermelho está uma tendência linear crescente e em azul uma tendência linear decrescente representadas em um plano cartesiano.

É dito linear, pois seu comportamento é equivalente a uma linha, onde a variação de um ponto ao outro é constante (sem alteração). Já crescente e descrescente se referem a como os valores de um eixo se comportam em relação ao outro: na tendência linear crescente, ao aumentarmos o valor em um eixo é esperado aumentarmos também o valor no outro eixo, já na tendência linear decrescente ocorre o inverso: ao aumentarmos o valor em um dos eixos, é esperado que o valor no outro eixo decaia de forma constante.

Na Figura 3.12 conseguimos visualizar o padrão exposto pela reta linear vermelha, ou seja, ao crescermos as notas em matemática, esperamos que cresça as notas em ciências naturais.

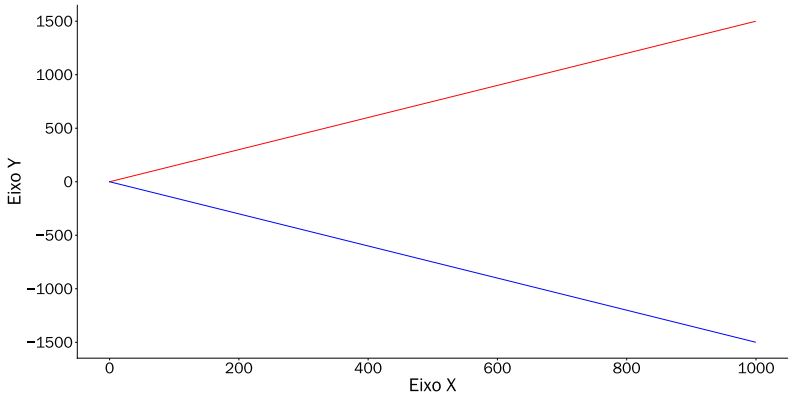

Figure 3.13: Tendências lineares em um plano cartesiano

Através da Figura 3.12 verificamos que, de certa forma, o tipo de escola que ele frequentou possui de fato impacto nas notas dos estudantes de Salvador, porém este padrão se avaliando outra prova?

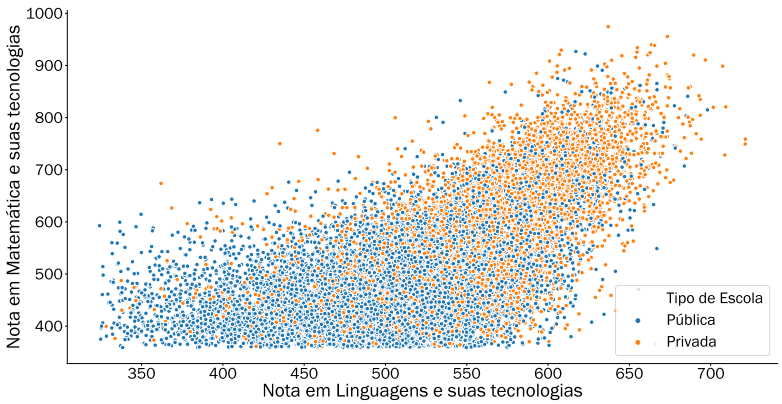

Figure 3.14: Relação entre nota de Linguagens e Matemática particionado pelo tipo de escola em 2019

A Figura 3.14 apresenta o gráfico de dispersão entre a nota em matemática (eixo Y) e a nota em Linguagens (eixo X) semelhante a Figura 3.12 e o padrão se repete: no geral, os estudantes de escolas públicas apresentam um rendimento inferior aos estudantes de escolas privadas.

Este conhecimento é importante para ressaltar a necessidade do aperfeiçoamento

das escolas públicas no município e buscar formas de reverter ou equiparar este quadro.

3.6 Histograma

Para expandir ainda mais as discussões referente ao quarto questionamento, vamos utilizar mais uma ferramenta gráfica de visualização: o histograma. Um histograma de um conjunto de dados numéricos se parece muito com um gráfico de barras apresentado anteriormente, embora tenha algumas diferenças importantes que examinaremos nesta seção.

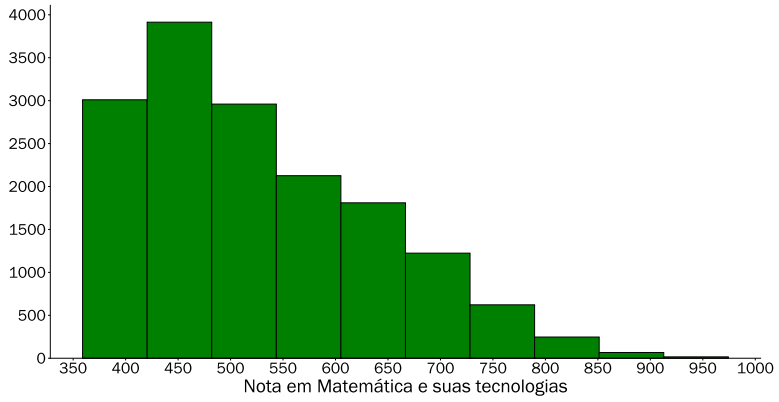

Figure 3.15: Histograma com 10 categorias de notas de matemática dos estudantes de escolas públicas e privadas de Salvador em 2019

A Figura 3.15 apresenta o histograma das notas dos estudantes de escolas públicas e privadas de Salvador em matemática no ano de 2019. No eixo horizontal está representado os valores numéricos das notas dos participantes agrupados em intervalos discretos. Fazendo um paralelo com o capítulo 2, o intervalo contínuo numérico estudado foi transformado para **K** valores categóricos/discretos. Este valor **K** é definido pelo usuário e na imagem anterior foi definido como igual a 10, ou seja, existem na Figura 3.15 dez categorias igualmente espaçada. Você pode perceber isso ao contar a quantidade de "caixinhas" que existem no histograma. Já o eixo vertical representa a quantidade de valores que estão em cada categoria, ou seja, quanto mais valores são representados por aquela classe maior será a altura de sua barra. Caso você esteja atento, provavelmente notou uma semelhança com a frequência absoluta apresentada durante a seção do gráfico de barras.

Antes de discutirmos o quarto questionamento, é importante entender que ao avaliar um histograma é preciso compreender que cada barra representa uma

categoria que define um intervalo numérico limitado. Esse intervalo é na maioria das vezes apresentado da seguinte forma:

$$[limite\ inferior,\ limite\ superior)$$

Onde o *limite inferior* representa o menor valor contido naquela categoria e *limite superior* o maior valor daquela categoria. Porém, na matemática os sinais [e) apresentam um significado específico, importantes para compreender a definição de uma categoria do histograma: o primeiro representa um intervalo fechado já o segundo um intervalo aberto.

Juntando todo este conhecimento é possível dizer que cada K categoria em um histograma contém seu limite inferior, mas não contém seu limite superior. Em outras palavras, uma determinada barra (categoria) não representa seu limite superior, logo uma categoria começa no limite inferior e termina no superior, sem inclui-lo.

Na 3.15 é possível observar que com 10 categorias, a medida que aumentamos a nota em matemática menos representativa se torna aquelas categorias, dificultando a visibilidade das barras. Além disso, o intervalo mais dominante se encontra na faixa entre 400 e 500 pontos. Caso seja desejado aumentar a resolução desses intervalos, será necessário realizar o aumento de categorias e isso pode ser conquistado ao aumentar o valor de K.

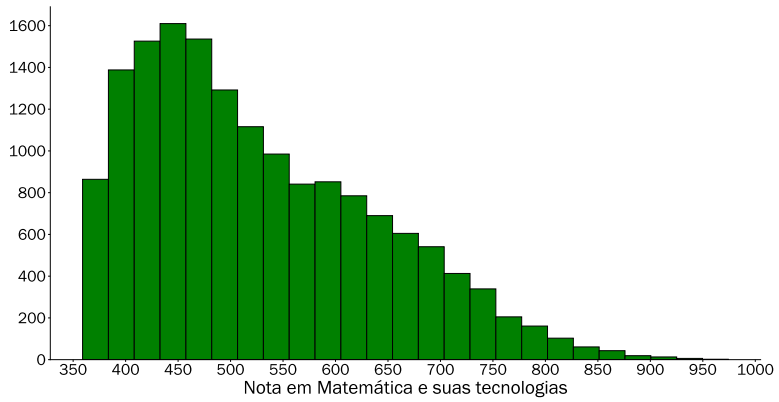

Figure 3.16: Histograma com 25 categorias de notas de matemática dos estudantes de escolas públicas e privadas de Salvador em 2019

Ao utilizar 25 categorias como apresentado na Figura 3.16 é possível identificar como mais precisão que o intervalo de notas mais dominante se encontra ainda mais próximo de 450 pontos. Porém este gráfico apresenta, sem distinção, estudantes de escolas privadas e públicas, mas separando, encontramos valores diferentes?

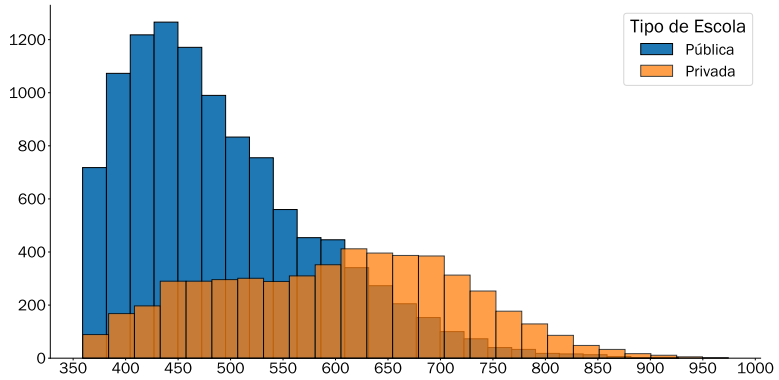

Figure 3.17: Histograma com 25 categorias de notas de matemática dos estudantes de Salvador em 2019 particionado pelo tipo de escola

A Figura 3.17 mostra o mesmo histograma agora com distinção entre os tipos de escola. Inicialmente verificamos que existe uma diferença na quantidade de estudantes de colégio público e privado na edição de 2019, como apontado na seção anterior. Além disso, é possível verificar que o perfil dos estudantes de colégio público é semelhante ao apresentado na Figura 3.16: O intervalo mais representativo está próximo de 450 pontos, porém para os estudantes de escola privada os intervalos mais representativos estão em entre 600 e 700 pontos.

Para uma melhor visualização e contornar o problema de diferença de grandezas, visto na seção 3.3, vamos separar os histogramas em diferentes gráficos com seus eixos representativos próprios:

A Figura 3.18 apresenta ambos os histogramas lado a lado com escalas de grandezas próprias. Você pode notar isso pelos valores máximos alcançados com a escola pública alcançando aproximadamente 1200 em uma categoria enquanto escola privada com valores próximos de 400 estudantes.

Ainda na Figura 3.18 é possível notar que os valores máximos, ou seja, os picos de cada histogramas são bem diferentes: para escola pública, o pico está entre 400 e 450 pontos enquanto para escolas privadas gira em torno de 600 a 650 pontos, aproximadamente 200 pontos de diferença. Além disso, um fator alarmante no histograma que representa os estudantes das escolas públicas é a queda nas notas de matemática a medida que a pontuação (eixo X) aumenta a partir dos 450 pontos. Este padrão também ocorre para as escolas privadas, porém para um valor superior a 700 pontos.

Assim, em relação a nota de matemática, podemos dizer que a resposta para o quarto questionamento: "**O tipo de escola (pública ou privada) pode**

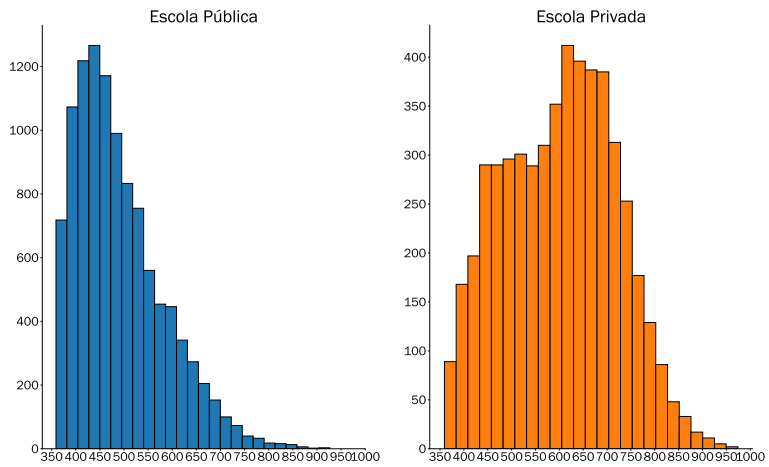

Figure 3.18: Histograma com 25 categorias de notas de matemática dos estudantes de Salvador em 2019 particionado pelo tipo de escola

influenciar nas notas dos estudantes neste exame?" apresenta uma resposta positiva, ou seja, é possível verificar uma diferença visual apresentada pelos histogramas e toda as discussões apresentadas até o momento. Fica a cargo do leitor avaliar se o comportamento das notas de matemática no ENEM 2019 se repetem para as outras avaliações do exame.

Nota: É importante ressaltar que alguns materiais trazem o conceito da densidade para o eixo vertical do histograma, porém dado o direcionamento do livro será mantido uma análise sem abordar este conceito dado a sua complexidade. A ideia de densidade é importante quando é analisado histogramas com intervalos de tamanhos diferentes, mas para intervalos iguais tanto o conceito de frequência absoluta (contagem) quanto densidade funcionam para o mesmo propósito.

Após concluir a leitura desta seção você pode notar a semelhança entre **histograma** e **gráfico de barras**, porém não confunda: eles são diferentes! Suas principais diferenças são:

- Os gráficos de barras exibem uma quantidade por categoria. Eles são frequentemente usados para exibir as distribuições de variáveis categóricas. Os histogramas exibem as distribuições de variáveis numéricas.
- Todas as barras em um gráfico de barras têm a mesma largura e há uma quantidade igual de espaço entre as barras consecutivas. As barras de um histograma podem ter larguras diferentes e são contíguas.

3.7 Concluindo ...

- Capítulo vai conter discussão final do objeto de estudo, resumo dos gráficos e discussão a respeito da importância da visualização gráfica (porém é necessário ter um ferramental estatístico para embasar as respostas)

3.8 Indo Além

- Avaliar o histograma para outras notas.