

Uma introdução gentil à Ciência de Dados

Kaike Wesley Reis, Júlia Bijos e Janaína Souza

2020-10-07

Conteúdo

| | |
|-------------------------------------------------------------------------------|-----------|
| Prefácio | 2 |
| 1 Introdução à Ciência de Dados | 3 |
| 1.1 O que são “dados” e onde estão presentes? | 6 |
| 1.2 O ciclo dos Dados - Construindo uma pergunta estatística | 9 |
| 1.3 Estruturando os dados | 12 |
| 1.4 Identificando o tipo de problema | 13 |
| 1.5 Considerações finais | 13 |
| 1.6 Referências | 14 |
| 2 Como organizar os dados | 16 |
| 3 Visualização e Ciência de dados | 17 |
| 3.1 Objeto de estudo do capítulo | 17 |
| 3.2 Gráfico de barra | 18 |
| 3.3 Gráfico de tendências | 22 |
| 3.4 Gráfico de setores | 26 |
| 3.5 Gráfico de dispersão | 30 |
| 3.6 Histograma | 33 |
| 3.7 Concluindo | 37 |
| 3.8 Indo Além | 38 |
| 3.9 Citações no capítulo | 39 |
| 4 Descrevendo e Construindo indicadores básicos com a ciência de dados | 40 |
| 4.1 Objeto de estudo | 41 |
| 4.2 Medidas de tendência central | 46 |
| 4.3 Medidas de dispersão | 46 |
| 4.4 Visualizando com <i>Boxplot</i> | 46 |
| 4.5 Distribuição | 46 |

Prefácio

Escrever prefácio.

Capítulo 1

Introdução à Ciência de Dados

Provavelmente você deve estar pensando que não faz ideia do que seja Ciência de Dados, já que nunca teve contato com esta área da ciência. Mas, será mesmo? Ao longo deste capítulo vamos entender o que é a **Ciência de Dados** e refletir como ela está inserida no nosso dia a dia.

A Ciência de Dados pode ser definida como o campo do conhecimento que busca transformar dados em informações.

Mas, o que isto significa?

Observe a Figura 1.1 que trata algumas situações vivenciadas na nossa cidade ou comunidade, que precisam ser adaptadas para melhorar o bem estar dos cidadãos. Você consegue imaginar como a Ciência de Dados poderia ajudar a lidar com estas dificuldades? Verifique os setores observados, os questionamentos e possíveis soluções indicados nas caixas em azul.

Esta imagem é um bom referencial de algumas aplicações da Ciência de Dados para o bem da sociedade. Além disso, mostra como podemos partir de problemas e perguntas iniciais para realizar investigações. Observe o questionamento 1, em que os ônibus não passam no horário esperado. Uma das formas de solucionar este problema é através de uma estimativa da frequência ideal de ônibus neste ponto, com base em um tempo de espera que se imagina ser adequado. Então, para definir esta frequência é necessário saber:

- a **frequência dos ônibus** que ali passam para conhecer a situação atual
- a **quantidade de passageiros** que utiliza o transporte
- a **melhor rota** que deve ser percorrida, de forma que evite atrasos no translado

A definição destes tópicos é importante, pois eles indicam quais dados são

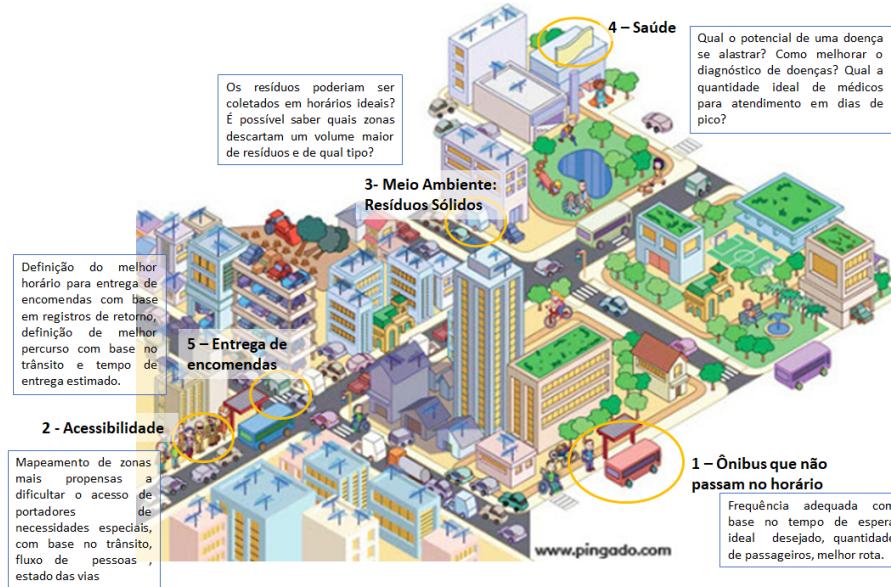


Figura 1.1: Ciéncia de Dados para o bem social

necessários obter. A partir do momento que são obtidos, o processo investigativo evolui até que se saiba qual é a frequência ideal de ônibus neste ponto da cidade!

Observe como temos questionamentos em tantas áreas diferentes. Isto indica que a Ciéncia de Dados é um campo do conhecimento que nos permite abordar problemas de uma forma abrangente e com aplicações em diversos setores!

No processo da Ciéncia de Dados, o **dado** é transformado em informação relevante por meio de etapas que permitem analisar tendências e prever comportamentos futuros! Estas informações geradas permitem extrair conclusões e criar **sacadas** (os famosos *insights* ou “lampejos de ideias”) para responder a perguntas e solucionar problemas.

Que tal conhecer alguns setores que tem aplicado Ciéncia de Dados aqui no Brasil? Transportes e Mobilidade Urbana: link 1 link 2

Saúde: link

Segurança pública: link

Comunicação com clientes: link

Turismo: link

Atividade Jurídica: link

Para aplicar esta ciéncia é preciso ter conhecimentos de Estatística, Computação e conhecimento sobre o problema investigado! Isto porque as ferramentas de solução são baseadas nestes conteúdos, por isso eles são a esséncia da Ciéncia de Dados! Mas não se engane, estas ferramentas “matemáticas” são vinculadas à ciéncias

sociais, biológicas, ambientais, ao setor de negócios, tecnologia, entre outros, a fim de descobrir padrões em problemas de diferentes naturezas (como vimos na Figura 1.1). Por este motivo, a Ciência de Dados é uma área **interdisciplinar**. A Figura 1.2 esquematiza os conteúdos básicos da Ciência de Dados.

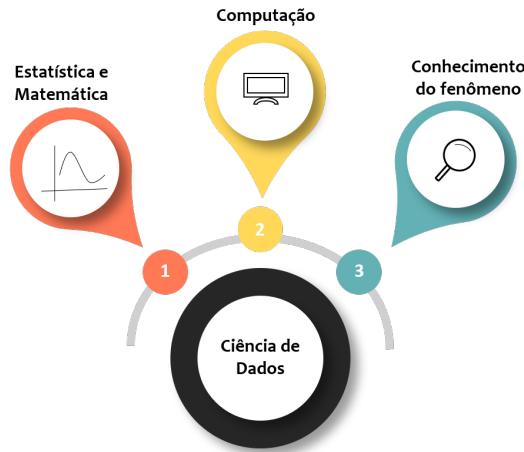


Figura 1.2: Interdisciplinaridade da Ciência de Dados

Mas por quê a Ciência de Dados se tornou indispensável?

A popularidade da área veio a partir do aumento de dados disponíveis atualmente. É impossível gerir tantos dados sem a utilização de computadores, que apresentam alta capacidade de processamento. Mas atenção, embora a Ciência de Dados seja favorecida pela tecnologia, é importante ressaltar que são seres humanos que direcionam, criam regras, avaliam e manuseiam todo o processo investigativo. Portanto, a criticidade de um profissional que avalie a execução de cada etapa realizada é essencial para garantir análises e interpretações coerentes a cada situação.

Para compreender melhor como o processo de Ciência de Dados ocorre, vamos pensar na seguinte situação:

Sabemos que atualmente o lazer está muito vinculado ao uso de tecnologias e, portanto, estamos a um clique de uma música que gostamos de ouvir, ou de um vídeo que queremos assistir, uma busca no *google* sobre algum tema de interesse. O fato de realizarmos estas buscas revela nossos interesses, você concorda? A partir da análise do nosso histórico de buscas, várias propagandas ou recomendações podem começar a nos ser feitas. Pense em quantas vezes você pesquisou sobre algum item e depois surgiram várias propagandas sobre ele. Ou quando você assistiu no *youtube* o clipe de uma banda e depois apareceram sugestões de outros clipes desta mesma banda, como na Figura 1.3. Observe que nesta imagem estamos assistindo a um vídeo da banda *Coldplay*, e ao lado existem várias sugestões de outras músicas deles, inclusive há uma indicação de outra banda.

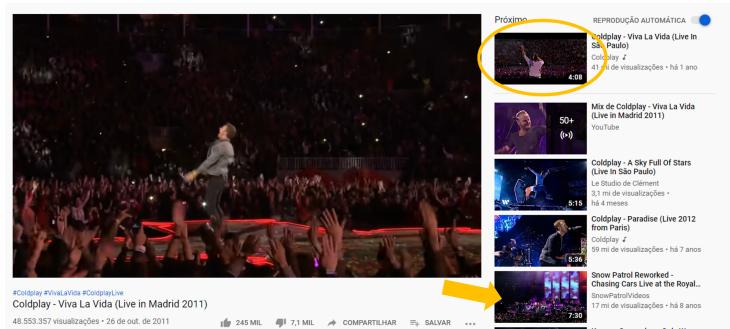


Figura 1.3: Mecanismo de recomendação do Youtube

Que tal entender melhor como este mecanismo ocorre?

Imagine que você será responsável por escolher um filme para assistir junto com seus amigos. Porém, eles disseram gostar de filmes de terror, romance, suspense, ação, comédia, ficção científica e drama. Para você ficou quase impossível escolher frente à tantas opções, já que quer ter certeza que eles irão curtir o filme. Assim, como podemos saber qual gênero de filmes deve ser escolhido? Sabendo isso, a sua escolha será certeira e a diversão estará garantida!

Podemos solucionar este problema utilizando a Ciência de Dados para analisar as preferências de filme dos seus amigos, de forma similar ao mecanismo de recomendações do *youtube*, por exemplo. Nas próximas seções deste capítulo, vamos solucionar cada etapa desta investigação!

Quer saber mais sobre onde a Ciência de Dados se aplica? Assista ao vídeo abaixo:

Depois de tantos exemplos uma conclusão é real: a Ciência de Dados está por toda parte e nós, fazemos parte dela ao consumir ou gerar dados. Você concorda?

1.1 O que são “dados” e onde estão presentes?

No item anterior vimos que a Ciência de Dados é algo indispensável, já que é impossível lidar com tantos dados sem o uso de tecnologias. Vimos também, que parte destes dados, somos nós quem geramos. Precisamos entender o que significa a palavra **dados** neste contexto.

Vamos dar sequência no nosso exemplo, onde queremos descobrir qual é o gênero de filmes que você deve escolher. Para isto, podemos avaliar qual é o gênero favorito de cada amigo seu a partir dos filmes que eles assistiram recentemente e o gênero destes filmes. Desta forma, iremos verificar se eles têm a preferência em comum por algum gênero. Para iniciar a investigação, devemos criar um registro para cada pessoa, contendo características importantes coletadas para a avaliação das preferências deles, como:

- Nome
- Filmes assistidos
- Gênero do filme

Portanto, o nosso registro irá conter observações de cada característica citada, para cada pessoa. Assim, dizemos que as observações destas determinadas características são os nossos dados.

Em outras palavras: dados são observações que foram coletadas e armazenadas de alguma forma. Inicialmente, compõem apenas registros e não apresentam relevância. Qualquer dado pode ser armazenado, caso contrário não pode ser considerado um dado.

O dado por si só não apresenta significado e por isso não serve para gerar respostas, interpretações e informações. Assim, somente após processar e transformá-los é que se torna possível tirar conclusões.

Parte do trabalho do investigador é avaliar quais dados são de fato importantes para o processo de análise. Muitas vezes temos uma grande quantidade de dados, mas ao avaliar a natureza do problema percebemos que nem todos são fatores importantes para a situação investigada.

No nosso exemplo, queremos saber qual gênero de filmes você deve escolher para assistir com seus amigos. Já vimos que algumas características são importantes para guiar a sua decisão final, mas podem existir outras que também complementariam nossos dados. Todavia, deve ser feita uma valiação sobre a importância delas para o problema abordado. Por exemplo, poderíamos coletar a altura e peso de cada amigo, mas, isso seria relevante para a nossa investigação? Claramente não, portanto não faria sentido registrar estes dados. Observe também que você obteve os dados por meio de uma pesquisa realizada com os seus amigos. Todavia, há muitas outras fontes de obtenção de dados. Basta lembrar que nós mesmos somos geramos dados quando interagimos em uma rede social.

Portanto, os dados podem ser obtidos pelo uso de celulares, computadores, sensores, registros escolares, pesquisas de opinião ou qualquer forma de registro. E porquê é tão importante entendermos a definição de dados e como eles são obtidos? Basicamente, porque eles são a essência da Ciência de Dados. Sem eles não é possível gerar informações e aplicar o processo investigativo. A análise dos dados permite observar uma **tendência** ou **padrão** em processos, fenômenos na natureza ou mesmo nos nossos comportamentos. E este é o grande objetivo da Ciência de Dados, **reconhecer padrões e interpretá-los** para tomar boas decisões!

1.1.1 Posso compartilhar dados?

Vamos voltar à nossa investigação sobre os filmes. Lembre-se que para nós é importante registrar os últimos filmes assistidos pelos nossos amigos, o gênero dos filmes e a identificação da pessoa. Podemos nos questionar se estes registros serão restritos à você que está analisando ou se serão abertos à qualquer pessoa

(inclusive seus amigos). É um questionamento pertinente? Haveria algum incômodo se qualquer pessoa tivesse acesso a estes registros?

Outro questionamento que poderia ser feito antes mesmo de seus amigos aceitarem participar do experimento é: como os dados serão utilizados e com qual finalidade?

Estas perguntas são importantes porque as informações adquiridas a partir dos dados revelam gostos pessoais e padrões de comportamento dos seus amigos. E, portanto, quem tiver acesso a estes dados vai ter conhecimento sobre as preferências deles. E a forma como esta informação será utilizada é extremamente importante. Assim, temos duas observações:

1. *dados são gerados a todo momento*
2. *dados são transformados em informações que revelam padrões desconhecidos.*

Por este motivo, empresas e organizações tem tanto interesse em deter dados de usuários dos seus serviços, pois isso permite conhecer o cliente a ponto de fazer ofertas que se adequem ao perfil de cada um. Mas, quais são as consequências dessa prática? Para compreender mais, vamos discutir sobre a **privacidade**.

1.1.2 Privacidade de dados

A privacidade antes de tudo é um direito. Este direito nos resguarda da exposição de nossas informações pessoais. O contexto atual de estarmos conectados, com uma constante troca de informação, traz algumas preocupações quanto à garantia da nossa privacidade.

Podemos começar citando o exemplo das publicações em redes sociais. Por meio delas, divulgamos sobre nosso local de trabalho ou estudo, quem são nossos familiares, nosso itinerário, datas importantes e tantas outras informações, na maioria das vezes sem refletir o que isto representa. E estes são os dados que nós sabemos que estamos divulgando!

Além disso, os aplicativos que temos em nossos *smartphones* podem ter acesso à nossa câmera, microfone e contatos. Sim, ao fazer o download de um aplicativo e concordarmos com os termos de condição de uso, damos acesso à todos estes dados. Você já leu os termos de condições antes de prosseguir com a instalação de um aplicativo?

Mas muito além do que publicamos, existe uma infinidade de dados que são coletados sobre nós que nem temos ideia. Eles alimentam grandes *bases de dados* de empresas, organizações ou instituições. Nossas preferências de lazer, política, estilo, gostos musicais, itens que compramos, informações bancárias, local de viagens, são convertidos em informações nas mãos de quem pode manipulá-los. Ficamos expostos, sendo influenciados por serviços e propagandas e, ao mesmo tempo, não temos acesso à forma que processam estes dados.

Assim, sempre que abrimos nossos aplicativos, automaticamente somos direcionados a interagir com posts de conteúdos preparados para prender a nossa atenção,

ou sempre existem propostas imperdíveis para adquirir itens que geralmente nos interessamos.

O ponto central que deve ser levantado aqui é que podemos sim fazer uso de aplicativos, redes sociais e sites, mas devemos ter criticidade para entender que somos monitorados e possivelmente influenciados. Para refletir mais sobre a privacidade dos dados, assista ao vídeo indicado.

*Já ouviu o termo **LGPD**?* Diante da problemática da privacidade e segurança dos dados, o Brasil aprovou a *Lei n° 13.709/18 (Lei de Proteção de Dados - LGPD)*, o que vai exigir a adequação de empresas e corporações que realizam coleta, tratamento, processamento ou comércio de dados em prol de garantir a privacidade e a segurança de usuários. Isso será feito por meio de políticas e planos de proteção de dados. Ao mesmo tempo, nós usuários deveremos estar mais atentos à segurança que as empresas oferecem aos nossos dados. (*Para saber mais sobre esta lei, assista ao vídeo*)

A rede social Facebook é um exemplo de organização que já iniciou as alterações recomendadas pela lei em busca de transparência. A Figura 1.4 exibe partes da mensagem que aparece ao realizar o acesso à página.



Figura 1.4: Notificação do Facebook

1.2 O ciclo dos Dados - Construindo uma pergunta estatística

No início do capítulo, definimos Ciéncia de Dados como um campo da ciéncia que realiza a transformação de dados em informação por meio de etapas. Neste

tópico, vamos compreender melhor sobre cada etapa que ocorre neste processo. Primeiramente, esta série de etapas é denominada como *Ciclo dos dados*.

A compreensão do ciclo dos dados dá uma noção geral sobre o que deverá ser realizado na metodologia de investigação, possibilitando um melhor planejamento de cada etapa.

Uma vez que a Ciência de Dados busca extrair padrões para lidar com problemas, é essencial que inicialmente se tenha uma pergunta a ser respondida. Esta pergunta irá direcionar todo o nosso processo em relação à quais dados devem ser coletados, quais são os melhores métodos de análise e qual a natureza do problema.

O Ciclo dos Dados compreende quatro etapas, como indicado na Figura 1.5:

O Ciclo de Dados

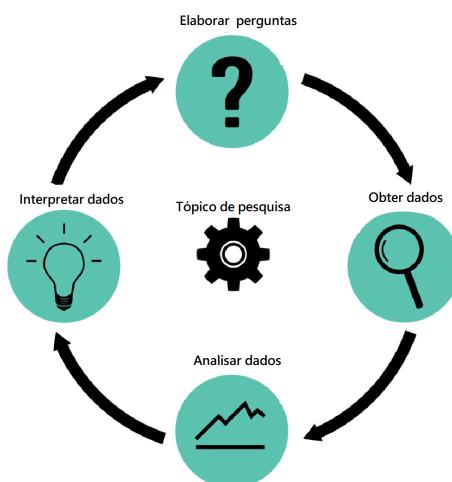


Figura 1.5: Etapas do ciclo de Dados

Vamos lembrar do nosso exemplo inicial, cuja pergunta é: Qual gênero de filme você deve escolher para assistir com seus amigos com base nas preferências deles?

Observe que geramos uma pergunta inicial que só poderá ser respondida a partir dos dados. Isto significa dizer que precisaremos coletar dados, analisá-los e, por fim, interpretá-los para tomar uma decisão. Por este motivo, esta pergunta é definida como **pergunta estatística**. A partir dela todas as outras etapas do ciclo dos dados serão direcionadas, a fim de respondê-la. A Figura 1.6 indica como cada etapa se desenvolve.

Mas, como saber se temos uma pergunta estatística ou não? Lembre-se que uma

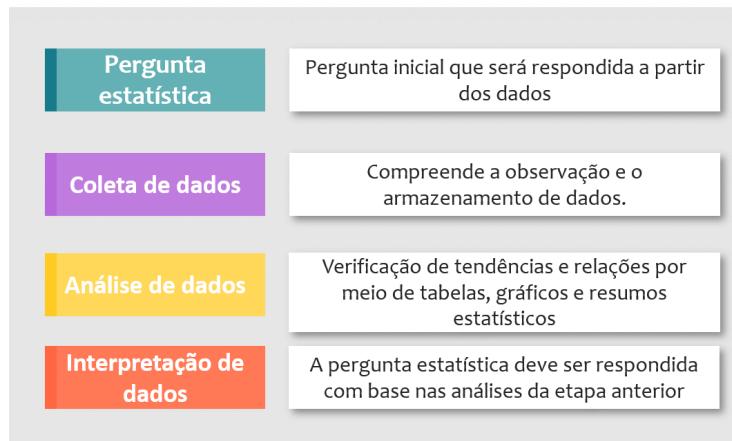


Figura 1.6: O que fazer em cada etapa?

pergunta estatística deve atender os requisitos citados no parágrafo anterior. Portanto, se eu te perguntasse: “Qual a sua idade?”, esta seria uma questão estatística?

Bom, você me responderia a sua idade. Porém, não seria necessário coletar mais dados para responder a pergunta, pois apenas uma única observação já foi suficiente. Ou seja, a etapa de análise de dados não se faz necessária e por isso, não chegamos à etapa de interpretação dos dados. Por estes motivos, comprovamos que esta questão não é uma pergunta estatística, pois não indica variabilidade.

Uma pergunta estatística sinaliza a variabilidade dos dados, que acontece quando existem observações que diferem da maioria registrada. Podemos adaptar a pergunta para que ela se torne uma questão estatística! No caso, poderíamos perguntar: “Qual é a idade dos estudantes do Projeto Ciência de Dados na Educação Pública?”. Ao coletar a idade de cada estudante perceberíamos que muitas se repetem, mas também há algumas que variam. Por exemplo, observamos estudantes de 11 a 15 anos. Todas as observações coletadas poderiam tranquilamente ser dispostas em um gráfico ou tabela que seriam usados para mostrar o padrão de idades da turma de estudantes do projeto. Graficamente, notamos que a maioria das idades equivale a 14 anos! Assim, todas as etapas do ciclo de dados se cumprem para responder esta pergunta.

Veja a Figura 1.7 que aponta diferenças entre estes dois tipos de perguntas.

Outro conceito relevante para qualquer investigação, é que para responder a uma pergunta estatística, podemos estabelecer **hipóteses**, que pode ser considerada como uma suposição que será testada ao longo do processo. Esta hipótese pode estar correta, neste caso ajudando a solucionar o problema, ou pode estar incorreta. Neste caso, precisamos investigar os motivos pelos quais ela é incorreta,

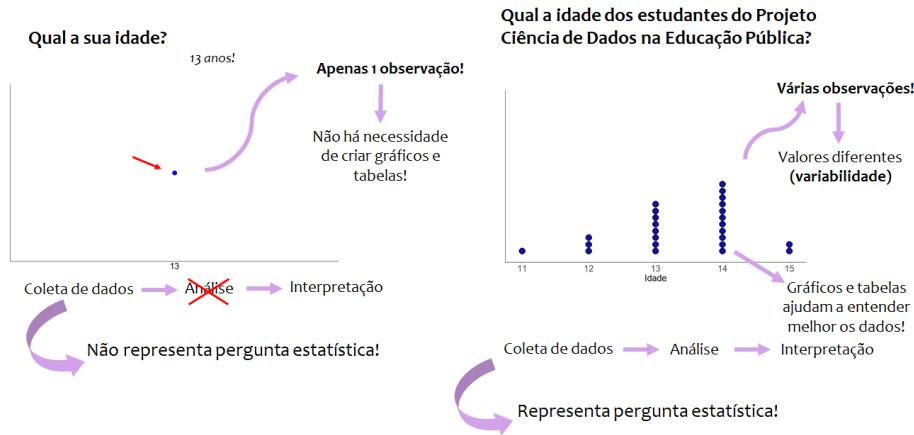


Figura 1.7: Pergunta estatística e não estatística

e pode até indicar que a nossa pergunta precisa ser melhorada. Pense no caso do nosso exemplo dos filmes, poderíamos estabelecer uma hipótese onde arriscamos dizer qual será o gênero preferido. Bom, esta hipótese será testada ao longo da investigação, e ao final, poderemos dizer se estava certa ou não.

1.3 Estruturando os dados

Após definir a pergunta estatística, devemos coletar os dados e armazená-los em algum formato. Chamamos de **estrutura de dados** o formato em que estes dados ficam armazenados.

A tabela é uma forma muito comum de se estruturar dados, embora não seja a única. Este formato é comum pois os dados ficam dispostos de uma forma organizada e de fácil entendimento.

A tabela é composta por linhas e colunas. Veja como exemplo a Tabela 1:

| Nome | Filme assistido | Gênero |
|-----------|------------------------------------------|---------|
| Gabrielle | Entre Realidades | Drama |
| Gabrielle | Getúlio | Drama |
| Gabrielle | Até que a sorte nos separe | Comédia |
| Gabrielle | Terremoto: A falha de <i>San Andreas</i> | Ação |
| Karen | A Lista de Schindler | Drama |
| Karen | Férias Frustradas | Comédia |
| Karen | Letra e Música | Romance |
| Karen | Pantera Negra | Ação |
| Isaac | Madagascar | Comédia |
| Isaac | Karatê Kid | Ação |

| Nome | Filme assistido | Gênero |
|-------|----------------------|---------|
| Isaac | Um senhor estagiário | Comédia |
| Isaac | A mulher de preto | Terror |

Observe que as colunas trazem as características dos dados que coletamos. Já as linhas trazem as observações coletadas para cada pessoa, a respeito das características. Perceba que para uma mesma característica temos observações que podem ser iguais ou diferentes. Embora um mesmo **Gênero** possa aparecer repetidas vezes, notamos que há observações que diferem. Por isso a tabela também permite enxergar a variabilidade dos dados. No capítulo 2 este tópico será abordado de forma mais aprofundada.

Além de representar os dados por meio de tabelas, você aprenderá no capítulo ?? como representar os dados de forma gráfica. Este formato permite visualizar as informações de uma forma mais clara e mais explicativa.

1.4 Identificando o tipo de problema

Um grande diferencial da Ciência de Dados é a investigação sobre o que os dados revelam acerca do futuro. Portanto, esta ciência não só obtém diagnósticos sobre situações já ocorridas como também traz *insights* sobre o que pode acontecer (Lembra deste termo? São aquelas **sacadas** que comentamos no início do capítulo). A isto chamamos de **predição**.

Parte do trabalho da Ciência de Dados é realizar previsões, e para isto, existem métodos estatísticos que podem ser aplicados. Em geral, podemos dividir as situações em problemas de **Regressão** ou **Classificação**. A Figura 1.8 exemplifica estes métodos.

Você irá aprender detalhadamente como aplicar estes métodos nos capítulos ?? e ?? deste *e-book*.

1.5 Considerações finais

Neste capítulo você foi apresentado à área de Ciência de Dados e percebeu como ela está presente no nosso dia a dia. Outro ponto relevante foi a percepção da nossa atividade enquanto consumidores e geradores de dados. Esta nova forma de gerar informações exige um conhecimento mínimo sobre como podemos ser influenciados a todo tempo.

Vamos finalizar o nosso exemplo?

Na tabela que construímos, temos 4 observações para cada pessoa. Ao analisarmos quantas vezes cada gênero aparece, temos:

- Drama: 3 observações

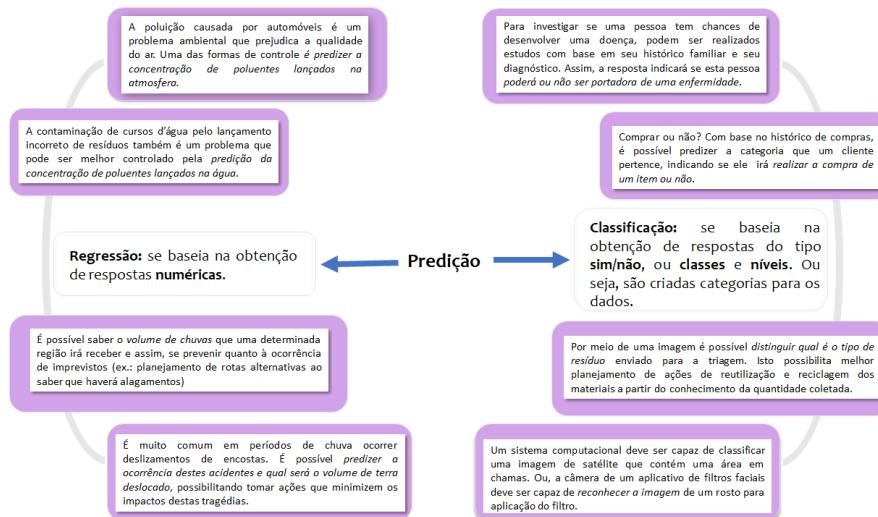


Figura 1.8: Regressão x Classificação

- Comédia: 4 observações
- Ação: 3 observações
- Romance: 1 observação
- Terror: 1 observação

Ao analisar o grupo de observações percebemos que o gênero **Comédia** aparece mais vezes no nosso conjunto de dados. Ao verificar a tabela, é possível notar que ele aparece pelo menos uma vez para cada pessoa. Portanto, é uma preferência comum a todo o grupo. Por isso, se você escolher um filme deste gênero sua chance de acerto será alta, concorda?

A Figura 1.9 sintetiza os conceitos discutidos neste capítulo introdutório.

Viu quantos conteúdos novos você aprendeu neste capítulo? Este aprendizado vai se aprofundar mais à medida que você avançar no estudo deste *e-book* e tiver curiosidade em relação aos assuntos abordados! A Ciéncia de Dados tem revolucionado os setores onde é aplicada, pois busca constantemente obter respostas valiosas. Portanto, o cientista de dados é movido pela curiosidade!

1.6 Referências

Tecmundo (2018). Do futebol à medicina: a ciéncia de dados está em todo lugar. Disponível em: <https://youtu.be/WjSimFnfPF0>.

Provocações Filosóficas (2018). A privacidade na internet. Diponível em: https://youtu.be/qw_TGrpPdkw.



Figura 1.9: Esquema de conceitos

Figura 1.1: Pingado sociedade ilustrativa (2010). Adaptada de: Cidade Sustentável. Disponível em: http://www.pingado.com/imagem/0811cidade_gra.jpg.

Figura 1.2: Coldplay (2011). Adaptada de: Coldplay - Viva la Vida (Live in Madrid 2011). Disponível em: <https://youtu.be/9ldOuVuas1c>.

LGPD Brasil. Lei Geral de Proteção de Dados - Lei nº 13.709/18. Disponível em: <https://www.lgpdblasil.com.br/>.

Brasil. LEI Nº 13.709, DE 14 DE AGOSTO DE 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.html.

McAfee (2019). Desafios Tecnológicos para atender a LGPD. Disponível em: <https://youtu.be/fuuudzh1qEo>.

Figura 1.5: O Ciclo de Dados. Adaptada de: Introduction to Data Science v_5.0 (IDS). Lesson 4: The Data Cycle. The Data Cycle file (LMR_1.3_Data Cycle)

Capítulo 2

Como organizar os dados

teste XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Capítulo 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 quanto 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 ciência de dados para **descrever um determinado processo ou objeto de estudo através de padrões e características que o representam**. Para este caso em específico, vamos analisar 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 acesso¹.

Como o termo **perfil** pode ser bem vasto e diversas características podem ser

extraídas do nosso objeto de estudo, é necessário concentrar essa análise em perguntas mais específicas para nortear o caminho. No decorrer deste capítulo, serão exploradas graficamente as seguintes questões:

- A quantidade de estudantes que realizaram 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 à 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, às 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.

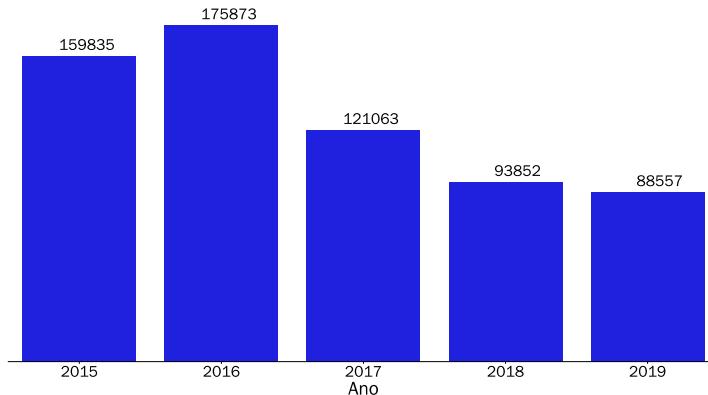


Figura 3.1: Quantidade de estudantes que se inscreveram o ENEM na capital bahiana

Na Figura 3.1 é apresentada a quantidade de estudantes, uma grandeza numérica, que realizaram o ENEM de 2015 até 2019 na capital bahiana. É possível notar

uma queda 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 realizaram 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, observa-se uma queda do número de inscritos no ENEM de Salvador, chegando a diminuir pela metade este número de 2016 para 2019.

Essa resposta pode levar a novos questionamentos, por exemplo, “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 padrões:

- 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 dada a quantidade finita de opções apresentadas. Essa pergunta está bastante relacionada 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:

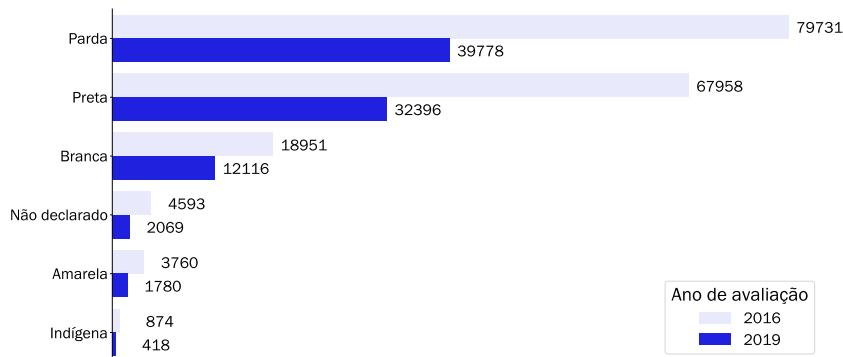


Figura 3.2: Distinção de estudantes inscritos por cor/raça da cidade de Salvador para os anos de 2016 e 2019

Através da Figura 3.2, são apresentados 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/amarelas que realizaram 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 2017², 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 definida 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:

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

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 de estudantes que realizaram em 2016. Transcrevendo a fórmula da frequência relativa apresentada anteriormente, temos:

$$\text{Frequência Relativa (\%)} = 100 * \frac{\text{estudantes que realizaram o ENEM em 2019}}{\text{estudantes que realizaram o ENEM em 2016}}$$

Como nos é apresentada 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 auto-declarados de cor branca todas as outras raças apresentaram uma proporção de aproximadamente 50%, ou seja, o número de estudantes pardos, pretos, amarelos,

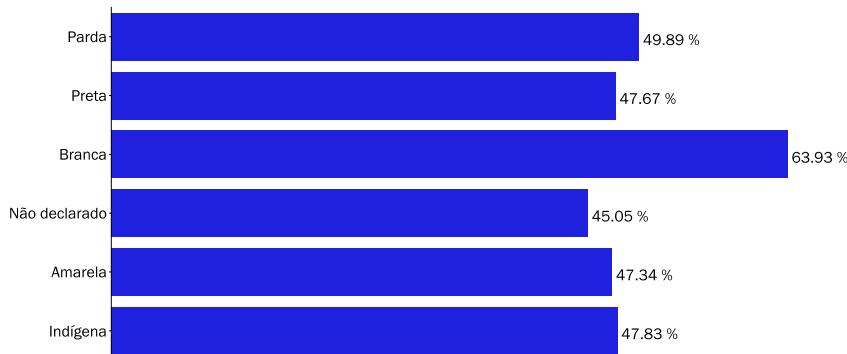


Figura 3.3: Comparação entre os estudantes inscritos de Salvador por cor/raça para 2016 e 2019

indígenas e não declarados caíram pela metade em comparação ao ano de 2016. Esta informação intensifica ainda mais o resultado apresentado na Figura 3.1, mostrando que ocorreu uma grande queda na quantidade de inscrições no geral, porém isso é verificado com maior intensidade entre estudantes não declarados de cor branca na capital bahiana ...

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 explorá-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 realizaram 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 ordem fundamental: importa saber se determinada observação ocorreu antes ou depois de determinado evento.

Este conceito será importante para expandir as análises realizadas apenas com os anos de 2016 e 2019 para a participação dos estudantes de Salvador por cor e raça apresentadas através dos gráficos de barras na seção 3.2. Será através deste tipo de gráfico que podemos avaliar como essa quantidade de inscrições se comportou (aumentou ou diminuiu) de 2015 até 2019 por raça, acompanhando sua tendência.

Note que realizamos este mesmo conceito no início da seção 3.2 demonstrando o número absoluto de inscrições no ENEM na capital bahiana de 2015 até 2019, porém quando vamos avaliar vários anos e possibilidades de raça/cor a utilização do gráficos de barras não demonstra ser a melhor opção, pois a visualização se torna muito carregada (cheio de elementos na tela).

Antes de mergulhar na análise desses dois questionamentos utilizando o gráfico de tendências é importante explorar mais um conceito novo: o plano cartesiano.

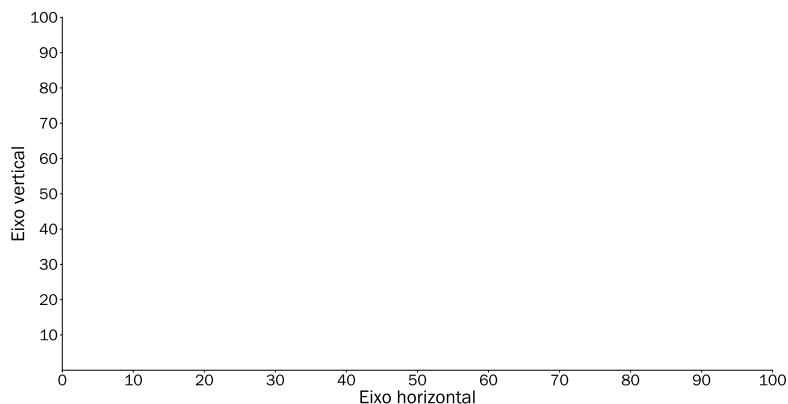


Figura 3.4: Plano cartesiano simplificado

A Figura 3.4 apresenta um plano cartesiano simplificado. São definidos dois eixos principais sendo eles o eixo horizontal e o eixo vertical. Cada eixo pode demonstrar o comportamento de uma variável desejada: Para o eixo horizontal, ao aumentarmos o valor se move para a direita e ao diminuir o valor se move para a esquerda, já para o eixo vertical, 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 observada 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

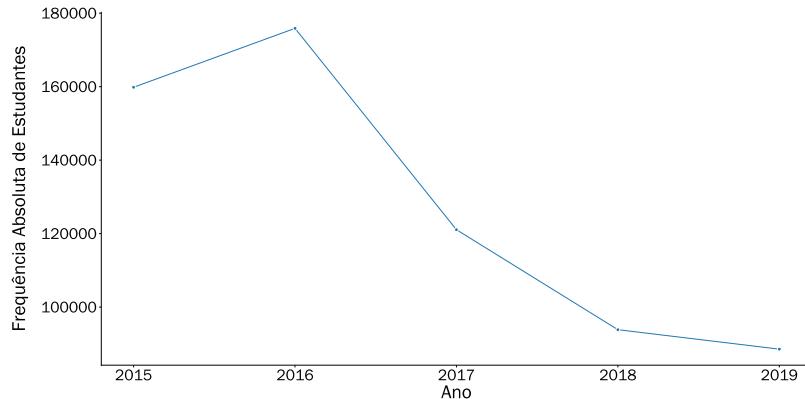


Figura 3.5: Quantidade de estudantes inscritos no ENEM na capital bahiana

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 pode ser definido como o maior valor identificado em um determinado período. 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.

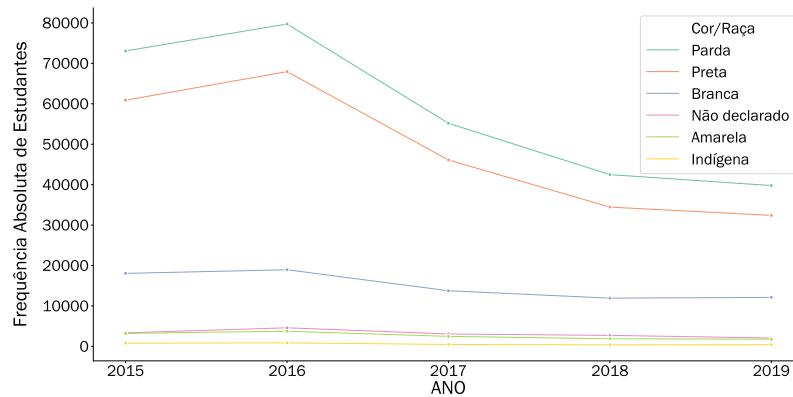


Figura 3.6: Tendência da quantidade de estudantes inscritos no ENEM por cor de 2015 até 2019

A Figura 3.6 é endereçada ao segundo questionamento. Podemos notar que

a tendência das duas primeiras curvas, referente as cores parda e preta dos 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 ao decorrer do tempo. 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 questionamento, cada curva foi separada de acordo a raça que ela representa:

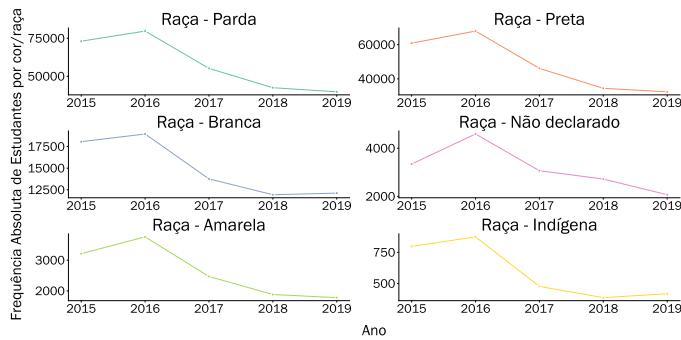


Figura 3.7: Tendência da quantidade de estudantes inscritos no ENEM participado por cor de 2015 até 2019

Na Figura 3.7 conseguimos notar a diferença de grandezas que foi mencionado anteriormente ao visualizar o eixo vertical (frequência absoluta de estudantes): para a cor parda, por exemplo, é possível enxergar valores próximos de setenta e cinco mil estudantes enquanto para a cor amarela os valores ficam próximos de três mil 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 números de inscrições apenas caem. Todavia, para as pessoas de cor branca e indígena o padrão se mantém, porém difere em 2019 onde ocorre um leve aumento em comparação ao ano anterior. Esse aumento no entanto é bem diferente ao considerar a ordem de grandeza entre as raças: enquanto para cor branca esse valor aumenta em torno de doze mil estudantes, para os indígenas eles aumentam em torno de 400 estudantes, ou seja, por mais que ambas as inscrições tenham aumentado, o número de inscritos de cor branca é

aproximadamente 30 vezes maior que o número de inscritos indígenas.

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 à educação. Além disso, é importante ressaltar a importância deste exame para conseguir ingressar nas faculdades ou universidades da cidade ou país, onde em Salvador infelizmente é mostrado uma tendência de saída dos estudantes nesse exame, principalmente aqueles de cor/raça negra e parda. Essa situação apresenta ainda mais a importância de integrar esses indivíduos para compreender a causa/motivo dessa evasão na capital bahiana.

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 com 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 à 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 a 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 à 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 apresentada a frequência relativa dos estudantes com e sem acesso à 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 tínhamos 72,6% estudantes com acesso à internet e esse valor aumentou para 84,7% em 2019, mostrando uma melhora de 12,1%. Essa melhora é mostrada visualmente através do tamanho da fatia referente a resposta “Sim” de 2015 para 2019. Este resultado pode estar associado a diversos fatores como maior acessibilidade a esta ferramenta como a redução de custos, aperfeiçoamento dos projetos sociais de inclusão digital e

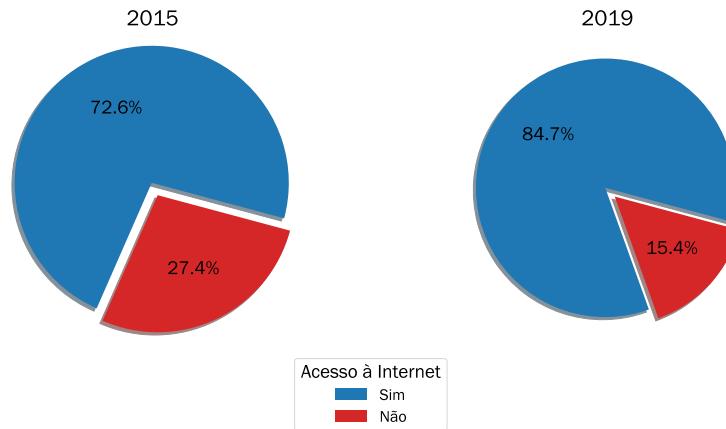


Figura 3.8: Porcentagem de estudantes inscritos no ENEM em Salvador com acesso a internet em 2015 e 2019

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 possuíam 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, associado ao encontrado na Figura 3.8 pode indicar que o acesso a internet realizado pelos estudantes podem surgir de outra fonte: o telefone celular, visto a queda considerável no acesso a computadores pessoais durante o mesmo período.

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

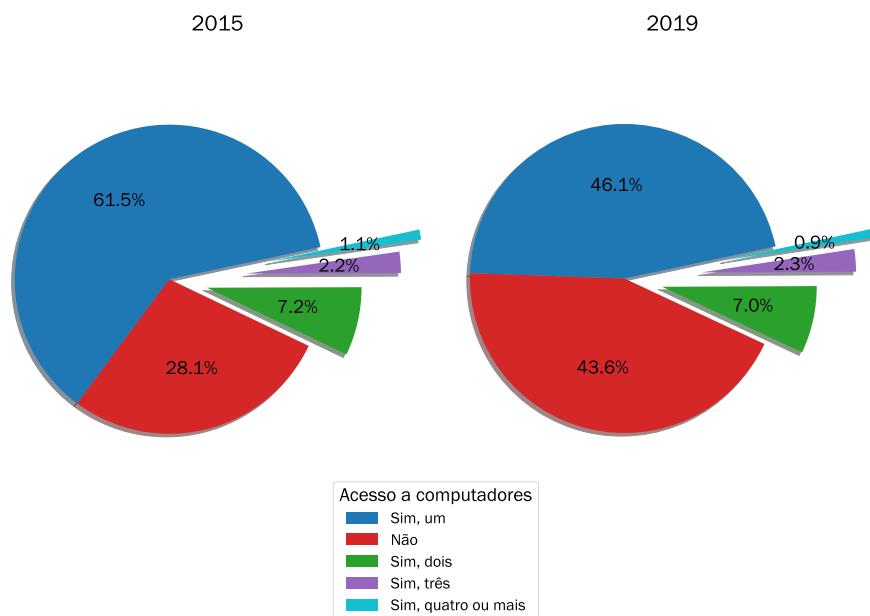


Figura 3.9: Porcentagem de estudantes inscritos no ENEM em Salvador com acesso a computadores pessoais em 2015 e 2019

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?

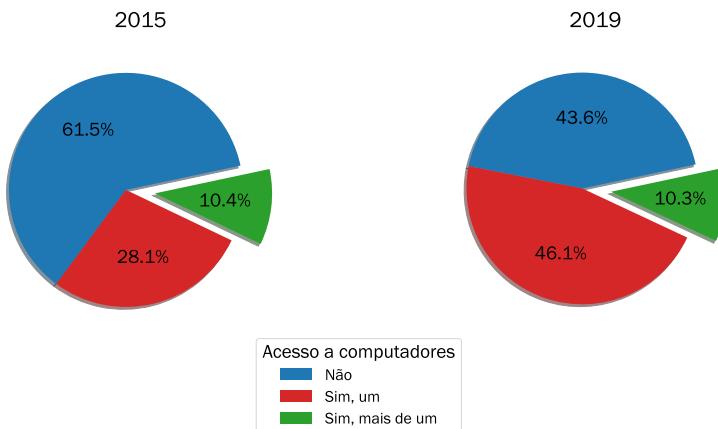


Figura 3.10: Porcentagem de estudantes inscritos no ENEM em Salvador com acesso a computadores pessoais em 2015 e 2019

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 e entender quem são as pessoas que sofrem deste tipo de necessidade digital.

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.

Na Figura 3.11 são mostradas as proporções de estudantes de Salvador que não possuem 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

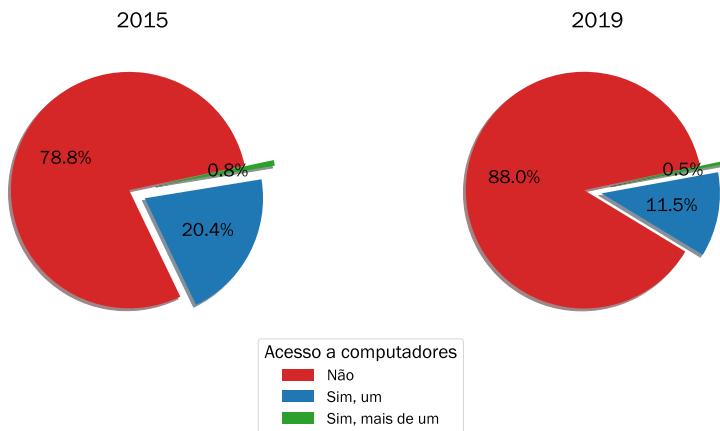


Figura 3.11: Porcentagem de estudantes inscritos no ENEM em Salvador sem acesso à internet em relação ao acesso a computadores pessoais em 2015 e 2019

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 e uso desta ferramenta para o seu aprendizado.

3.5 Gráfico de dispersão

Até o momento conseguimos observar os dados e refletir sobre três dos quatro 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 da escolha desta ferramenta 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 atípicos nas análises, pois apenas pessoas ausentes no exame possuem suas notas zeradas (com exceção da nota em Redação) e para focar nossa análise em estudantes de escolas públicas e privadas, desconsiderando aqueles que optaram por não informar o tipo de colégio. 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 na capital bahiana, ou seja, 15996 estudantes no total sendo 10760 de escola pública e 5236 de escola privada.

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.

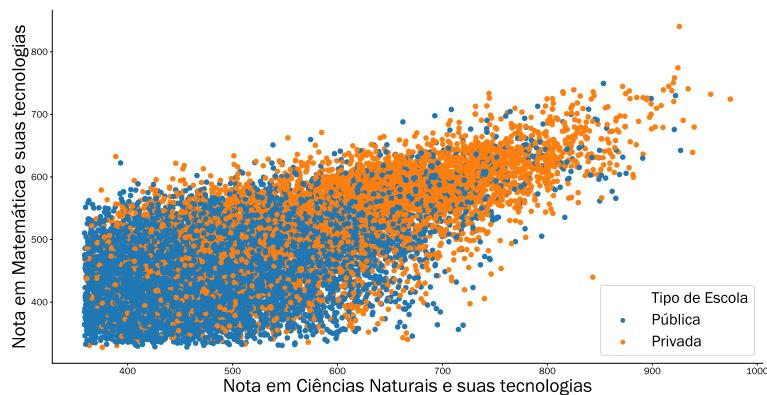


Figura 3.12: Relação entre nota de Ciências Naturais e Matemática particionado pelo tipo de escola em 2019

Na Figura 3.12 são apresentadas as notas dos estudantes de Salvador em matemática no eixo vertical e no eixo horizontal as notas em ciências naturais, destacando em cores o tipo de colégio: azul escola pública e em amarelo escola privada totalizando assim três variáveis representadas em uma só imagem. Neste gráfico de dis-

persão 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 (*nota em ciências naturais, 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 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 essas tendências são as mais comuns e intuitivas na natureza.

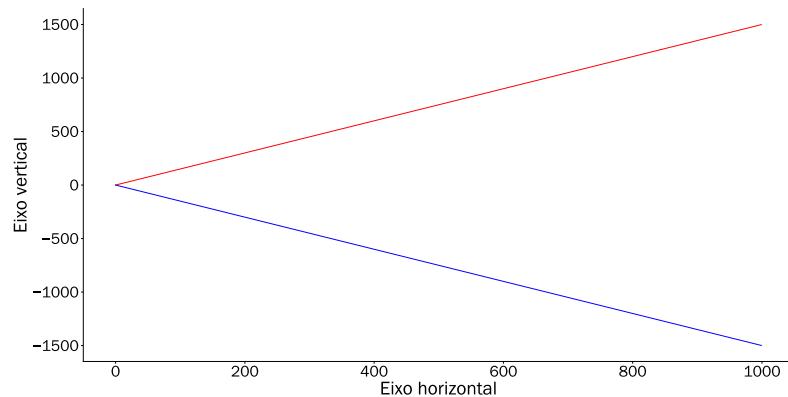


Figura 3.13: Tendências lineares em um plano cartesiano

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 em um dos eixos em relação a outro é constante (sem alteração).

Os termos 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 crescemos as notas em matemática, esperamos que cresça as notas em ciências naturais.

Através da Figura 3.12 verificamos que, de certa forma, o tipo de escola que o estudante frequentou possui um fato impacto nas notas dos estudantes de Salvador, porém este padrão se repete caso seja avaliado outra prova?

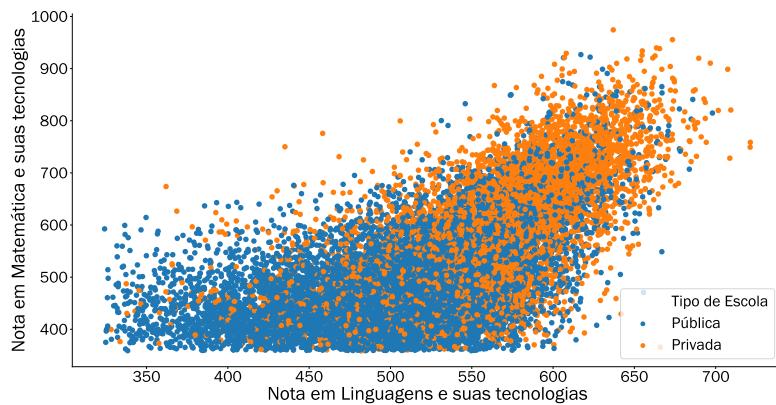


Figura 3.14: Relação entre nota de Linguagens e Matemática particionado pelo tipo de escola em 2019

A Figura 3.14 apresenta o gráfico de dispersão entre a nota em matemática (eixo vertical) e a nota em Linguagens (eixo horizontal) 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 que impacta de forma negativa padrões e classes sociais, dificultando o ingresso de estudantes de escolas públicas em cursos mais concorridos como Engenharia, Direito e Medicina.

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.

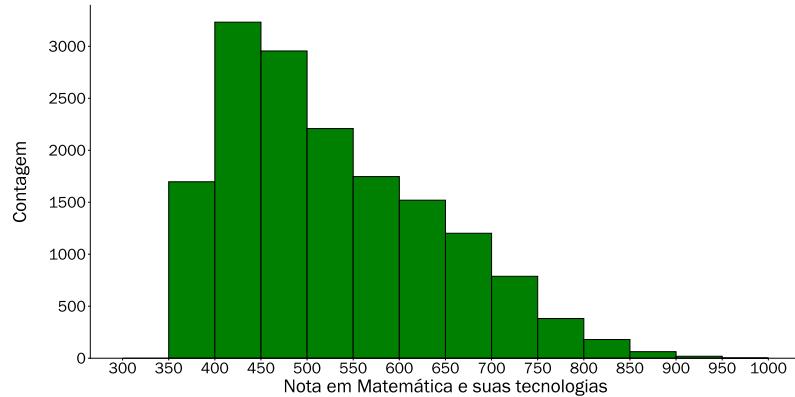


Figura 3.15: Histograma com as notas de matemática dos estudantes de escolas públicas e privadas de Salvador em 2019 de 300 à 1000 pontos com resolução de 50 pontos

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** pode ser definido pelo usuário de duas formas: um valor inteiro, onde o algoritmo irá particionar os números em **K** categorias de mesmo tamanho (largura), ou através de intervalos definidos pelo próprio usuário como foi feito na Figura 3.15 onde foram definidos limites de 50 em 50 pontos começando em 300 pontos até 1000 pontos. 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:

$$(\text{limite inferior}, \text{limite superior})$$

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 incluí-lo.

Na 3.15 é possível observar que 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 450 pontos. Caso seja desejado melhorar a resolução desses intervalos, será necessário realizar o aumento de categorias.

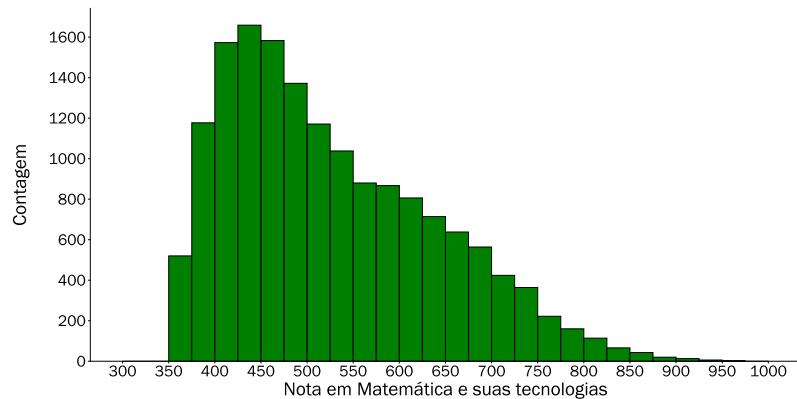


Figura 3.16: Histograma com as notas de matemática dos estudantes de escolas públicas e privadas de Salvador em 2019 de 300 à 1000 pontos com resolução de 25 pontos

Ao utilizar um espaçamento de 25 pontos como apresentado na Figura 3.16 é possível identificar como mais precisão os intervalos de notas dos estudantes da cidade de Salvador. Podemos destacar agora, através da Figura 3.16 que o intervalo mais representado é aquele que começa em 425 pontos até 450 pontos. Isso foi possível graças ao aumento da quantidade de categorias por meio da diminuição do espaçamento de 50 pontos para 25 pontos.

Porém este gráfico apresenta, sem distinção, estudantes de escolas privadas e públicas, mas separando encontramos valores diferentes?

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

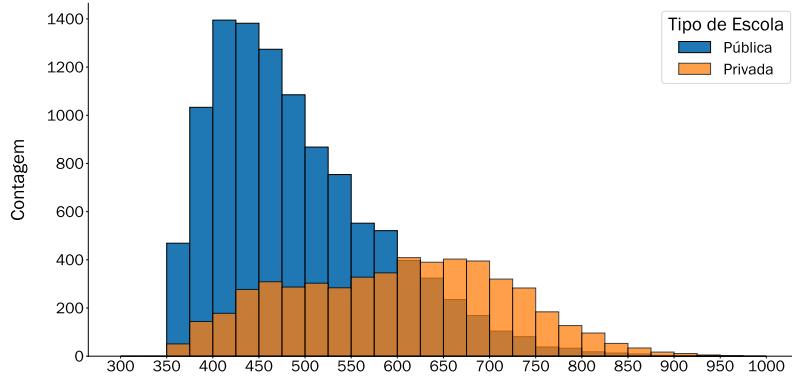


Figura 3.17: Histograma com 25 categorias de notas de matemática dos estudantes de Salvador em 2019 particionado pelo tipo de escola

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 escala entre os tipos de escolas, visto anteriormente na seção 3.3, vamos separar os histogramas em diferentes gráficos com seus eixos representativos próprios:

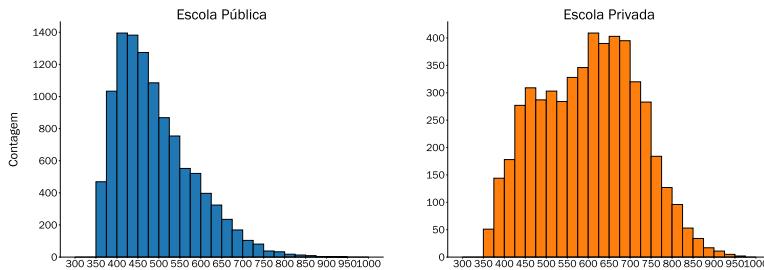


Figura 3.18: Histograma com 25 categorias de notas de matemática dos estudantes de Salvador em 2019 particionado pelo tipo de escola

A Figura 3.18 apresenta ambos os histogramas lado a lado com escalas de grandeza próprias. Você pode notar isso pelos valores máximos alcançados onde na escola pública foi alcançado aproximadamente 1400 em uma categoria enquanto na escola privada foi alcançado 400 em uma categoria.

Ainda na Figura 3.18 é possível notar que os valores máximos de cada histogramas são bem diferentes: para escola pública, as categorias mais representativas começam em 400 pontos e vão até 475 pontos, já para as escolas privadas gira em torno de 600 a 700 pontos, apresentando assim uma diferença em torno de 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 horizontal) aumenta a partir dos 475 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 influenciar nas notas dos estudantes neste exame?**” apresenta uma resposta positiva, ou seja, é possível verificar uma diferença visual entre os tipos de escolas através dos histogramas discussões 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 visto 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 ...

Através deste capítulo conseguimos entender como a visualização gráfica pode trazer diferentes ideias e esclarecimentos a respeito dos nossos questionamentos, apresentando informações de forma simples e de fácil entendimento. Vimos também que cada gráfico pode trazer uma visão distinta e cabe ao leitor saber escolher qual a melhor abordagem a partir da sua pergunta e conjuntos de dados. Nosso questionamento sobre o **perfil dos estudantes de Salvador que realizaram o ENEM** conseguiu apresentar diversos *insights*, porém desanimadores

...



Figura 3.19: Infográfico dos resultados encontrados para o nosso questionamento

Na Figura 3.19 é apresentado o infográfico resumindo as informações extraídas a partir da análise gráfica. Esses resultados são desanimadores, pois no geral mostra uma grande evasão no número de inscritos de estudantes de vulnerabilidade social e um baixo rendimento dos estudantes de escolas públicas em comparação aos de escola privada. Todavia, para confirmar esses resultados apenas a visualização gráfica é insuficiente: Na ciência de dados precisamos de indicadores e medidas matemáticas para expressar se de nossas hipóteses são verídicas. Esse ferramental será explorado nos próximos capítulos deste livro, então mantenha o estudo!

3.8 Indo Além

Fizemos diversas análises e respondemos alguns questionamentos a respeito do **perfil dos estudantes de Salvador que realizaram o ENEM** não foi? Porém com a riqueza que esta base de dados possui o leitor pode explorar ainda mais!

Utilizando Python, explicada em nosso capítulo ?? de programação, você conseguiria **ir além** e responder os seguintes questionamentos?

- Através do infográfico na Figura 3.19 avaliamos que ocorreu um aumento no número de estudantes sem acesso a computadores pessoais no ENEM de 2015 para 2019, o que pode dificultar os estudos desta parcela de estudantes. Você consegue avaliar a distribuição de cor/raça para esses estudantes em 2019 utilizando o gráfico de barras?
- Na seção 3.6 verificamos que, infelizmente, na edição de 2019 do ENEM as notas dos estudantes de escolas públicas são menores em comparação aos

estudantes de escolas privadas para a prova de Matemática. Utilizando o histograma, você consegue avaliar se este padrão se repete para as notas de Linguagens nesta mesma edição?

3.9 Citações no capítulo

- [1] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Microdados Exame Nacional do Ensino Médio.** Disponível em: [link de acesso](#)
- [2] Acorda Cidade. **IBGE-BA: Salvador é a capital mais negra do Brasil e com a maior desigualdade salarial entre brancos e pretos.** Publicado em 19 de novembro de 2018. Disponível em: [link de acesso](#)

```
##  
## Attaching package: 'dplyr'  
## The following object is masked from 'package:kableExtra':  
##  
##     group_rows  
## The following objects are masked from 'package:stats':  
##  
##     filter, lag  
## The following objects are masked from 'package:base':  
##  
##     intersect, setdiff, setequal, union
```

Capítulo 4

Descrevendo e Construindo indicadores básicos com a ciência de dados

Apesar da ciência de dados ser bem conhecida atualmente através das aplicações mais práticas e incríveis da Inteligência Artificial como demonstrado no vídeo abaixo:

INserir VÍDEO DO YOUTUBE AQUI

A ciência de dados está presente de outras formas no dia a dia do brasileiro de forma sutil.

Na Figura 4.1 em destaque conseguiram identificar esses termos? Eles estão muito presentes neste imenso universo que é a ciência de dados e seus conceitos são bem simples de entender!

Porém, você pode estar se questionando “Porque devo aprender mais sobre eles?” ou “Vou usar isso em algum momento da minha vida?”. Questões como podem ser recorrentes durante a aprendizagem e são importante serem endereçadas. A área que é apresentada na Figura 4.1 e será discutida neste capítulo é chamada de **estatística descritiva**. Ao estudar os dados, é comum o cientista buscar padrões desconhecidos e quantificar grandes quantidades de números em um só valor e nesse aspecto que esta ferramenta simples e eficaz é aplicada. O nome pode parecer complicado, porém se trata de um ramo da matemática com uma série de técnicas para descrever e resumir uma grande massa de informações. Essa área pode ser dividida em três grupos de medidas: tendência central, dispersão e forma. Neste capítulo focaremos nas duas primeiras. Além disso iremos explorar formas de visualizações utilizando estes conceitos como o diagrama de caixas e distribuições.

CAPÍTULO 4. DESCREVENDO E CONSTRUINDO INDICADORES BÁSICOS COM A CIÊNCIA DE DADOS

Primavera começa na segunda com temperaturas acima da média e pouca chuva na maior parte do país, diz Inmet

Os termômetros deverão registrar temperaturas de 1°C a 1,5°C acima da média. Chuvas deverão se concentrar no RS e SC no início da estação.

19/09/2019 às 08:52 - por G1

Notas médias do Enem 2019 caem em todas as provas objetivas

Inep divulgou nesta sexta-feira (17) o resultado do Enem. As notas médias em todas as disciplinas caíram. 53 participantes tiraram nota 1 mil na redação.

Por G1

17/01/2020 10h33 · Atualizado há 6 meses

Vitória da Conquista tem o janeiro mais seco dos últimos 80 anos

Amplitude térmica atinge cidade do sudoeste da Bahia.

Estação de meteorologia da Uesb não esperava grande variação.

Do G1 BA

Figura 4.1: Manchetes de jornais com termos da ciência de dados

4.1 Objeto de estudo

Para compreender a importância dessas medidas estatísticas e como usa-las vamos estudar os dados de segurança pública da cidade de Salvador disponibilizados pela Secretaria de Segurança Pública (SSP) no portal para compreender um pouco da realidade que Salvador convive: a violência. Porém, antes de apresentar nosso tema central de estudo precisamos entender essa base de informações.

Disponibilizados através de boletins mensais, as ocorrências dos principais delitos na capital bahiana são separados por áreas e regiões. Os principais tipos de delitos considerados são:

- Homicídio Doloso
- Lesão Corporal Seguida de Morte
- Roubo com Resultado Morte (Latrocínio)
- Tentativa de Homicídio
- Estupro
- Roubo a Ônibus (Urbano e em Rodovia)
- Roubo de Veículo

CAPÍTULO 4. DESCREVENDO E CONSTRUINDO INDICADORES BÁSICOS COM A CIÊNCIA DE DADOS

- Furto de Veículo
- Uso/Porte de Substância Entorpecentes (Usuários)

Você viu que eu citei “áreas” e “regiões” certo? Elas são definidas pela SSP em Salvador respectivamente como Área Integrada de Segurança Pública(AISP) e Região Integrada de Segurança Pública(RISP). Para entender melhor essas divisões vamos usar uma abordagem de conjuntos:

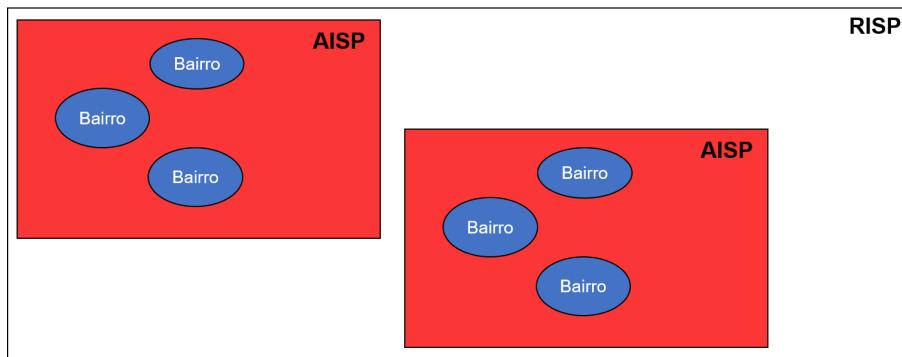


Figura 4.2: Conceito por trás das divisões AISP e RISP

Através da Figura 4.2 podemos verificar que a abordagem de conjuntos explica bem os conceitos: **AISP** são conjuntos de bairros, ou seja, cada **AISP** representa uma quantidade de bairros da cidade. Já **RISP** se trata de um conjunto de **AISP** e finalmente, o conjunto total de **RISP** representa toda a extensão da cidade de Salvador. Essa divisão é realizada para facilitar a vida dos profissionais de segurança em compreender como cada região se comporta de acordo a um determinado tipo de delito, além de agrupar melhor os bairros, que somam um valor superior a 160 em Salvador. Imagina como seria mais complexo de apresentar um plano para cada bairro em específico?

As Figuras 4.3, 4.4 e 4.5 mostram as divisões referentes as AISP e RISP em Salvador. Para contextualizar, a maioria das escolas que participaram do projeto “Meninas na Ciência de Dados” são localizadas no bairro da federação e este bairro está contido na AISP - Rio Vermelho que, por sua vez, está contida na RISP - Atlântico.

Com toda estas informações em mente como tipos dos principais delitos, divisões e subdivisões territoriais de Salvador determinada pela SSP podemos de fato identificar um objeto de estudo com um propósito: entender um pouco a violência em Salvador. Será analisado em específico um tipo de delito principal nos meses de janeiro, fevereiro e março de 2019 em Salvador: **Roubo a Ônibus urbano e em rodovia**. Essas escolhas não foram aleatórias. Os três meses citados foram escolhidos por ser um período de grande movimentação na capital: estação de verão, principal estação para turismo em Salvador, e conhecido por ser um período onde a grande maioria dos trabalhadores entram de férias. Além disso o

CAPÍTULO 4. DESCREVENDO E CONSTRUINDO INDICADORES BÁSICOS COM A CIÊNCIA DE DADOS

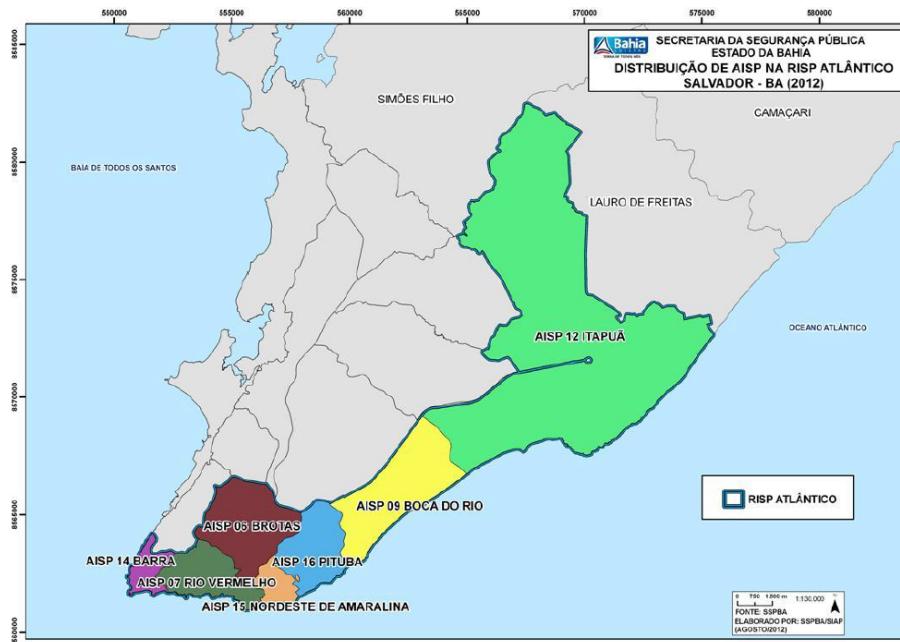


Figura 4.3: RISP Atlântico de Salvador realizado pela SSP

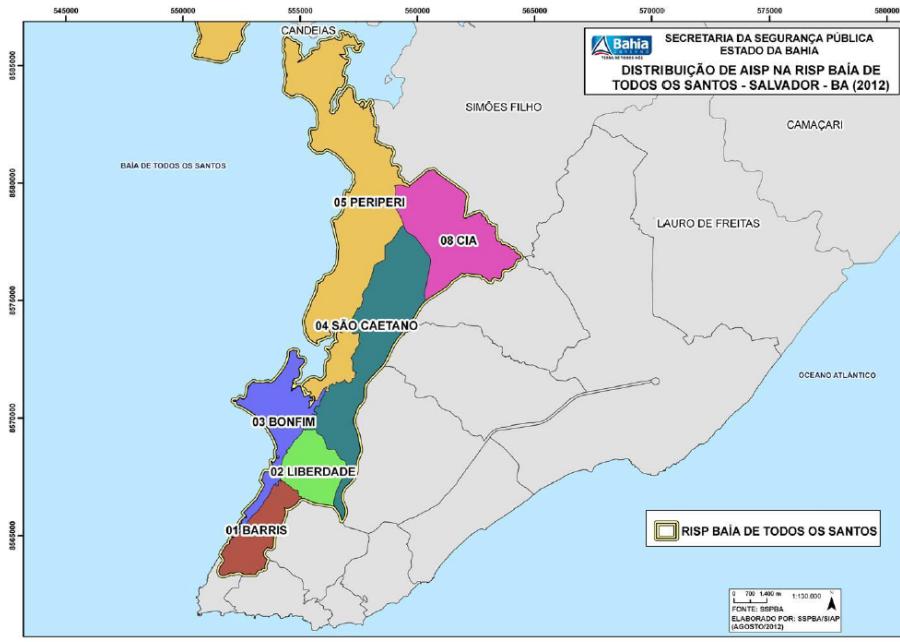


Figura 4.4: RISP Baía de Todos os Santos (BTS) de Salvador realizado pela SSP

CAPÍTULO 4. DESCREVENDO E CONSTRUINDO INDICADORES BÁSICOS COM A CIÊNCIA DE DADOS

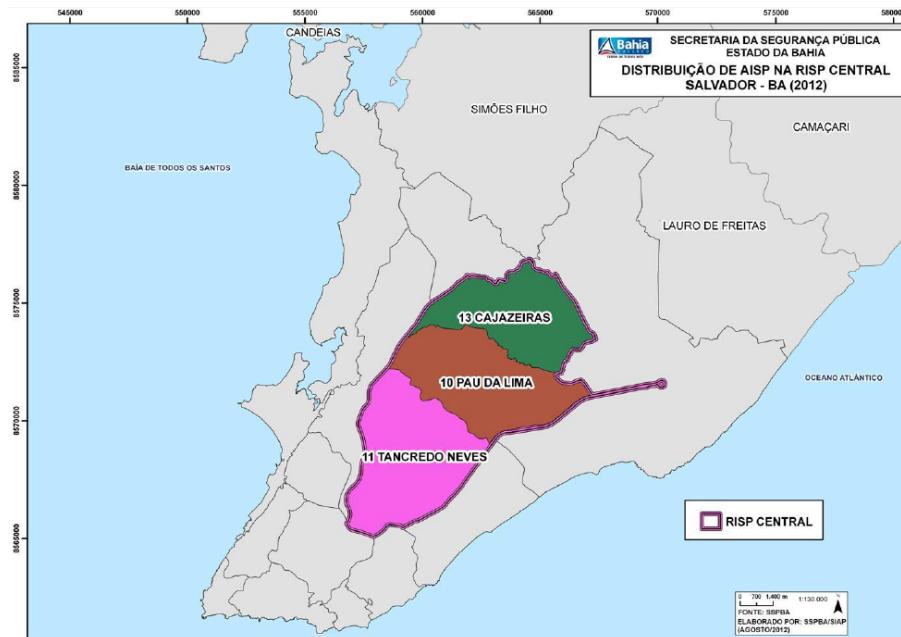


Figura 4.5: RISP Central de Salvador realizado pela SSP

delito de roubo a ônibus é uma realidade presente para quem depende do serviço público para se locomover na capital bahiana, seja a trabalho ou a lazer.

Na Figura 4.6 podemos verificar por exemplo que **em média ocorreu 3 assaltos em Salvador por dia avaliados durante um ano** e neste capítulo frases como essa serão compreendidas por vocês de forma mais simples e intuitiva!

A base de dados referente aos assaltos à coletivos na cidade de Salvador na estação do verão em 2019 é apresentada abaixo:

CAPÍTULO 4. DESCREVENDO E CONSTRUINDO INDICADORES BÁSICOS COM A CIÊNCIA DE DADOS

Salvador registra quase dois mil assaltos a ônibus em 2019

Dados da SSP-BA ainda apontam que de janeiro a dezembro deste ano, capital baiana teve uma média de três assaltos por dia.

Por Giana Matiazz, TV Bahia e G1 BA
17/01/2020 21h21 - Atualizado há 8 meses

Passageiros de ônibus são assaltados em diferentes pontos de Salvador: 'disse que ia dar tiro'

Três ônibus foram alvos de assaltantes na manhã desta quarta-feira (26). Testemunhas contam que uma passageira foi agredida em um dos coletivos.

Por Tv Bahia e G1 BA
26/08/2020 07h54 - Atualizado há um mês

Figura 4.6: Manchetes do jornal G1 sobre roubos a ônibus na cidade de Salvador em diferentes períodos

| RISP | AISP | Janeiro | Fevereiro | Março |
|-----------|----------------|---------|-----------|-------|
| Atlântico | Brotas | 17 | 6 | 16 |
| Atlântico | Rio Vermelho | 7 | 10 | 5 |
| Atlântico | Boca do Rio | 5 | 1 | 7 |
| Atlântico | Itapuã | 17 | 11 | 8 |
| Atlântico | Barra | 2 | 2 | 1 |
| Atlântico | Nordeste | 4 | 1 | 6 |
| Atlântico | Pituba | 13 | 8 | 4 |
| BTS | Barris | 17 | 5 | 10 |
| BTS | Liberdade | 8 | 8 | 11 |
| BTS | Bonfim | 8 | 12 | 12 |
| BTS | São Caetano | 22 | 15 | 16 |
| BTS | Periperi | 13 | 11 | 20 |
| BTS | CIA | 1 | 1 | 2 |
| Central | Pau da Lima | 5 | 4 | 9 |
| Central | Tancredo Neves | 40 | 25 | 33 |
| Central | Cajazeiras | 11 | 9 | 6 |

Nesta tabela vemos valores para cada uma das regiões divididas pela Secretaria de Segurança Pública que serão estudados no decorrer deste capítulo. Alguns valores são bem alarmantes como a AISP de Tancredo Neves com um total de 40 ocorrências de assalto à coletivos em janeiro e 33 em março. Em contrapartida, outras AISP como Barra e CIA apresentam valores muito baixos em comparação com 1 ou 2 ocorrências.

Porém, e se quisermos resumir esses valores para gerar indicadores para

CAPÍTULO 4. DESCREVENDO E CONSTRUINDO INDICADORES BÁSICOS COM A CIÊNCIA DE DADOS

uma determinada região (RISP) ou para a cidade nesta época de verão? Como esses indicadores poderiam ajudar os gestores a entender como está determinada região ou área em relação ao aumento da violência, considerando este delito? E principalmente: Como podemos visualizar e passar estas ideias de forma fácil e intuitiva para gerar ações públicas de combate?

Para responder esses e outros questionamentos, vamos estudar alguns conceitos importantes da ciência de dados como medidas de tendência central, medidas de dispersão, diagrama de caixas e distribuição no decorrer deste capítulo.

4.2 Medidas de tendência central

4.3 Medidas de dispersão

4.4 Visualizando com *Boxplot*

4.5 Distribuição