

Introdução a Machine Learning

Módulo 1 - Manipulação e visualização de dados

Foto de La-Rel Easter,
disponível na Unsplash.
Adaptada pelo autor.

Foto de Jason Leung,
disponível na Unsplash.
Adaptada pelo autor.



Sumário

Aula 1 - A história dos computadores1

Aula 2 - Áreas de atuação e carreiras em ML4

2.1 Estatística6

2.2 Inteligência Artificial (IA)7

2.3 Aprendizado Profundo (AP)8

2.4 Ciência de Dados (CD)9

Aula 3 - Introdução a dados 10

3.1 Tipos de dados 12

3.2 Dados Tabulares..... 13

3.3 Operações Comuns 14

3.4 Pipeline 15

3.5 Coleta de Dados 16

3.6 Informações dos dados..... 17

3.7 Pré-processamento dos dados 18

Aula 4 - Ética e sensibilidade dos dados 19

Aula 5 - Visualização de dados21

5.1 Visualizações fundamentais 23

Explore mais!28

Referências Bibliográficas.....29

Foto disponível no FreePik.
Adaptada pelo autor.

Foto de Sebastian Dumitru,
disponível na Unsplash.
Adaptada pelo autor.

Aula 1 - A história dos computadores

Com o avanço dos computadores, cada vez mais surgem possibilidades para o seu uso com o objetivo de solucionar problemas do dia a dia. Vemos essa tendência em empresas e até mesmo na sociedade em geral, visto que a cada dia surgem novas tecnologias, porém, nem todas as coisas que vemos divulgadas hoje no mercado são realmente novas. Muitas das tecnologias que vemos hoje em dia já apareceram em filmes ou livros de ficção científica e apareceram pela primeira vez em pesquisas a décadas atrás.

Para entender um pouco melhor sobre o **Aprendizado de Máquina**, veremos a seguir qual sua origem a partir da historia do computador e como essa tecnologia amplamente utilizada vem sendo aplicada no mundo da tecnologia atualmente.



A palavra “computador” vem de muito antes da criação das máquinas que carregam esse nome hoje em dia. Essa palavra vem do latim “computare”, onde “com” significa “junto” e “putare” significa “calcular, avaliar, estimar”.



A palavra computador, antigamente, era utilizada para se referir às pessoas que computavam e, um pouco antes dos computadores que conhecemos hoje ficarem populares, as pessoas que faziam cálculos eram chamadas de "computadores". A NASA, por volta do ano de 1950, tinha uma sala com vários computadores, na sua maioria mulheres, que passavam o dia calculando a trajetória de foguetes.

Em 1936, o matemático Alan Turing (figura 1), propôs um modelo teórico de computador programável capaz de simular qualquer forma de computação algorítmica. Esse modelo ficou conhecido como **Máquina de Turing**. Durante a Segunda Guerra Mundial ele foi convidado junto a um grupo pequeno de mentes brilhantes, para desvendar as cartas com mensagens escondidas que os alemães usavam para se comunicar. Neste período, utilizando recursos do governo britânico, construiu uma máquina que se enquadrava ao modelo que ele propôs anos atrás. Essa máquina ficou conhecida como ***Electronic Numerical Integrator and Computer (ENIAC)***, em português, Computador e Integrador Eletrônico Numérico, e o ano de 1946 ficou conhecido como o ano em que o primeiro computador surgiu.

Figura 1 – Alan Turing, matemático inglês considerado o pai da computação¹.

¹Fonte: Revista Galileu. Disponível em: <https://revistagalileu.globo.com/Cultura/noticia/2018/06/17-fatos-e-curiosidades-sobre-vida-do-alan-turing.html>. Acesso em: agosto de 2023.



Foto disponível no
FreePik. Adaptada
pelo autor.

Foto de National Cancer Institute, disponível
na Unsplash. Adaptada pelo autor.

Aula 2 - Áreas de atuação e carreiras em ML

O aprendizado de máquina é uma **técnica sofisticada para resolver problemas computacionais**. Normalmente, a solução de problemas utilizando o computador é dada por um passo a passo, uma receita, de como o computador deve fazer para achar a solução. Porém, existem problemas que são complexos demais para serem resolvidos dessa maneira.

As técnicas de ML envolvem um conjunto de dados e um modelo:

- **Modelo:** funciona como se fosse uma base de como o computador vai fazer para aprender;
- **Dados:** funcionam para ensinar o modelo sobre o assunto em questão.

Podemos fazer uma analogia com a sala de aula, onde o "modelo" seria responder perguntas e os "dados" seriam as perguntas em si. Assim, podemos ter perguntas de matemática, se quisermos aprender matemática, ou perguntas de português, se quisermos aprender português. Nesse caso, os dados mudam, mas o "modelo" de aprendizado segue o mesmo.

A área de Aprendizado de Máquina é uma área em constante expansão, e existem diversas carreiras relacionadas a ela. Os profissionais podem atuar em várias carreiras, como, por exemplo:

- Desenvolvedor de *software* para *Machine Learning*;
- Cientista de dados;
- Engenheiro de *Machine Learning*;
- Especialistas em IA (Inteligência Artificial);
- Arquiteto de dados;
- Analista de dados;
- *Data engineers*.

Além disso, há também a possibilidade de atuar em empresas de diversos setores, como tecnologia, finanças, saúde, varejo, entre outras. É importante ressaltar que a área de aprendizado de máquina é multidisciplinar, envolvendo conhecimentos em matemática, estatística, ciência da computação e áreas específicas de aplicação.



2.1 Estatística

A estatística é uma área que utiliza a matemática para coletar, interpretar e analisar dados de pesquisas sobre a natureza, sociedade, economia, mercado e muito mais. Ela teve origem quando os governos se interessaram em obter informações quantitativas e qualitativas sobre suas riquezas, tributos, populações e moradias. O primeiro censo foi publicado em 1662 e muitos historiadores consideram esse marco como a origem da estatística, porém, no Egito antigo os faraós já ordenaram registros de dados sobre as suas colheitas.

Os modelos de aprendizado de máquina foram desenvolvidos com base nos modelos estatísticos comentados anteriormente, muitos deles com o mesmo funcionamento descrito de anos atrás, porém agora com o auxílio dos computadores para fazerem a parte difícil.

Dentro da estatística diversos modelos são utilizados para facilitar essas análises. Esses modelos são construções de hipóteses a partir da análise de dados, de sua relação e de outras variáveis para prever ou comprovar fatores. Eles funcionam como se fosse um esqueleto que se adapta ao problema em questão para achar correlações e prever valores.

Um dos modelos estatísticos mais conhecidos é **regressão linear**, que foi inventada no ano de 1877, mas ainda é utilizada em análises hoje em dia.

Com a facilidade da comunicação de hoje, o acesso e a coleta de dados se tornaram muito mais simples, todavia, a quantidade de dados acaba se tornando um problema, **pois na internet muitos dados são gerados em apenas um minuto** e para que toda essa informação seja processada, é preciso utilizar os computadores. Eles têm a capacidade de trabalhar com volumes imensos de dados, tarefa que demoraria anos para um humano fazer.

2.2 Inteligência Artificial (IA)

A Inteligência Artificial ou no inglês, *Artificial Intelligence (IA)* é o ramo da computação que busca **criar máquinas capazes e imitar o comportamento humano**. Esse tipo de problema normalmente envolve uma solução utilizando modelos de Aprendizado de Máquina, ou Aprendizado Profundo, que serão explicados no futuro.

Existem diversos estudos que lidam com tarefas de inteligência artificial, como robôs conversacionais (*chatbots*) e reconhecimento de objetos/pessoas em imagens. Qualquer tarefa que busque fazer com que o robô se passe por um humano é considerada IA.

Hoje em dia, um robô que consiga parecer ser um humano, sem que as outras pessoas notem é uma realidade distante, mas isso não impede os filmes de ficção científica criarem várias teorias e cenários de como seria ter uma tecnologia assim.



Figura 2 – Cena do filme M3GAN, lançado em 2023, dirigido por Gerard Johnstone. M3GAN é uma boneca realista com inteligência artificial, programada para ser a melhor amiga de uma criança².

²Fonte: MCNAB, Kaitlyn. People think M3GAN looks like renesmee and more hilarious reactions to horror movie trailer. Teen Vogue, 2022. Disponível em: <https://www.teenvogue.com/story/people-think-m3gan-looks-like-twilight-renesmee-and-more-hilarious-reactions-to-horror-movie-trailer>. Acesso em: agosto de 2023.

2.3 Aprendizado Profundo (AP)

O Aprendizado Profundo, também conhecido como *Deep Learning (DL)*, é uma técnica mais recente e poderosa. Ela tem o intuito de **imitar o funcionamento do cérebro humano com os "neurônios" e "sinapses"**. Essa técnica teve sua primeira publicação no ano de 1943, três anos antes de Alan Turing construir o primeiro computador programável. Porém, ela só começou a ser utilizada na década de 1990, pois antes disso os computadores não tinham potência suficiente para lidar com a quantidade de dados necessária nessa técnica.

Essa forma de solucionar problemas é amplamente utilizada nos dias de hoje. Muitos dos problemas de IA são resolvidos com esse tipo de algoritmo. São algoritmos complexos, porém seu funcionamento se assemelha muito aos algoritmos de aprendizado de máquina, que são mais simples de trabalhar. Por conta disso, esta trilha abordará de maneira aprofundada apenas os modelos de ML e não os de DL.



Ilustração feita por rawpixel.
com, disponível no Freepik.
Adaptada pelo autor.

2.4 Ciência de Dados (CD)

Ciência de dados ou *Data Science (DS)* é uma área que envolve o uso de métodos de *Machine Learning* para extrair *insights* e conhecimento a partir de dados. Ela combina elementos de diferentes áreas, mas tem como objetivo **chegar a conclusões úteis para tomadas de decisão em diversas áreas**, tais como negócios, saúde, finanças, governo, entre outras.

Uma das principais características da ciência de dados é a sua capacidade de trabalhar com grandes volumes de dados. Por meio de técnicas de mineração de dados, análise exploratória e modelagem estatística, os cientistas de dados conseguem extrair informações valiosas a partir de dados brutos, muitas vezes em tempo real.

Essas informações podem ser utilizadas para gerar *insights* importantes e apoiar tomadas de decisão estratégicas em diferentes áreas e setores. Com o aumento da disponibilidade de dados e a evolução de tecnologias e algoritmos, a ciência de dados tem se tornado cada vez mais relevante e demandada nos mais diversos segmentos da economia.



Foto disponível no
FreePik. Adaptada
pelo autor.

Aula 3 - Introdução a dados



Foto de Jake B, disponível
na Unsplash. Adaptada
pelo autor.

Os dados são essenciais para a área de Aprendizado de Máquina, pois é através deles que algoritmos podem ser treinados e aprimorados. No geral, quanto mais dados relevantes e diversificados forem utilizados no processo de treinamento, maior será a capacidade do modelo em fazer previsões precisas e generalizar para novos dados. Além disso, a qualidade dos dados e o seu pré-processamento também são fatores críticos para o sucesso de um modelo de Aprendizado de Máquina. Por isso, é importante garantir a integridade, qualidade e diversidade dos dados!



É importante saber identificar corretamente o tipo de dado para aplicar as técnicas de pré-processamento, transformação e análise corretas.

Ilustração feita por upklyak, disponível no Freepik. Adaptada pelo autor.

3.1 Tipos de dados

Uma amostra de dados é um subconjunto de dados coletados que representa um universo maior de dados. O objetivo de coletar uma amostra é inferir informações sobre o universo maior de dados, sem ter que analisar todo o conjunto de dados, o que muitas vezes é inviável...

Os dados podem ser classificados em:

- **Dados numéricos:** são aqueles que podem ser quantificados em uma escala numérica, como altura, peso, temperatura, etc;
- **Dados categóricos:** são aqueles que não possuem uma ordem ou hierarquia natural, como cor, sexo, raça, etc;
- **Dados ordinais:** possuem uma ordem ou hierarquia natural, como nível educacional, posição em uma corrida, entre outros.

3.2 Dados Tabulares

Podemos considerar como dados tabulares informações organizadas em **tabelas com colunas e linhas, onde cada coluna representa uma variável e cada linha corresponde a uma observação, registro ou valor**. Os formatos mais comuns de tabela são o *CSV (comma-separated values)* e o Excel.

Entre as características dos dados tabulares, incluem-se:

- Tipo do dado (numérico, categórico, texto);
- Presença de valores ausentes;
- Distribuição dos valores.

Devido ao seu formato fácil de entender e manipular, a maioria dos dados disponíveis publicamente estão armazenados dessa forma.

3.3 Operações Comuns

As operações com tabelas são cruciais em todas as fases da análise de dados. A seguir, são resumidas algumas das operações mais comumente utilizadas durante o seu manuseamento.

- Geralmente, quando queremos **criar ou carregar *dataframes* (tabelas)** em *Python*, a biblioteca Pandas é a mais adotada. Utilizamos a função *read_csv* ou *read_excel* para isso;
- **Para fatiar dados**, é possível utilizar a notação de índices e colchetes do *Python*, assim como as funções *iloc* e *loc* do Pandas;
- A **seleção condicional** permite escolher apenas as linhas de uma tabela que atendem a uma determinada condição, como por exemplo, selecionar apenas os valores maiores do que um certo número;
- **Para adicionar e remover colunas** em uma tabela, basta utilizar a notação de colchetes e o operador de atribuição do *Python*;
- A operação de ***join/merge/concat/flip*** são usadas para juntar dois ou mais *dataframes* em um único ou transformar as colunas em linhas, e vice-versa;
- Por fim, **a operação de agrupamento (*group by*)** é usada para agrupar os dados de acordo com uma determinada coluna e realizar operações estatísticas sobre as colunas restantes, como média, soma, desvio padrão, entre outras.

3.4 Pipeline

O *Pipeline* de dados é composto por etapas que vão desde a coleta até a análise de dados. O processo começa com a definição do problema e dos objetivos do estudo, seguido pela coleta de dados de diversas fontes, como bancos de dados, dispositivos IoT, arquivos CSV, entre outros. É importante garantir que os dados coletados sejam confiáveis, completos e diversificados, para que possam representar de forma adequada o fenômeno estudado.

Em seguida, os dados são pré-processados, o que inclui diversos procedimentos, entre eles:

- Tratamento de valores ausentes;
- Normalização;
- Detecção de *outliers*;
- Redução de dimensionalidade.

Em seguida, os dados são transformados em um formato adequado para análise, geralmente em forma de tabela ou matriz, utilizando bibliotecas de programação como o Pandas, do *Python*.

A etapa seguinte é a visualização dos dados, que ajuda a entender melhor as relações entre as variáveis e a identificar padrões e tendências nos dados. Existem diversas ferramentas e bibliotecas disponíveis para criação de gráficos e visualizações, como *Matplotlib*, *Seaborn* e *Plotly*. Além disso, é fundamental entender os tipos de visualizações mais comuns e como aplicá-las.

Por fim, os dados são analisados e interpretados utilizando técnicas de aprendizado de máquina, estatística e outras abordagens. **O objetivo final é extrair insights e conhecimento a partir dos dados**, que possam ser aplicados em tomadas de decisão, otimização de processos ou desenvolvimento de novos produtos e serviços.

3.5 Coleta de Dados

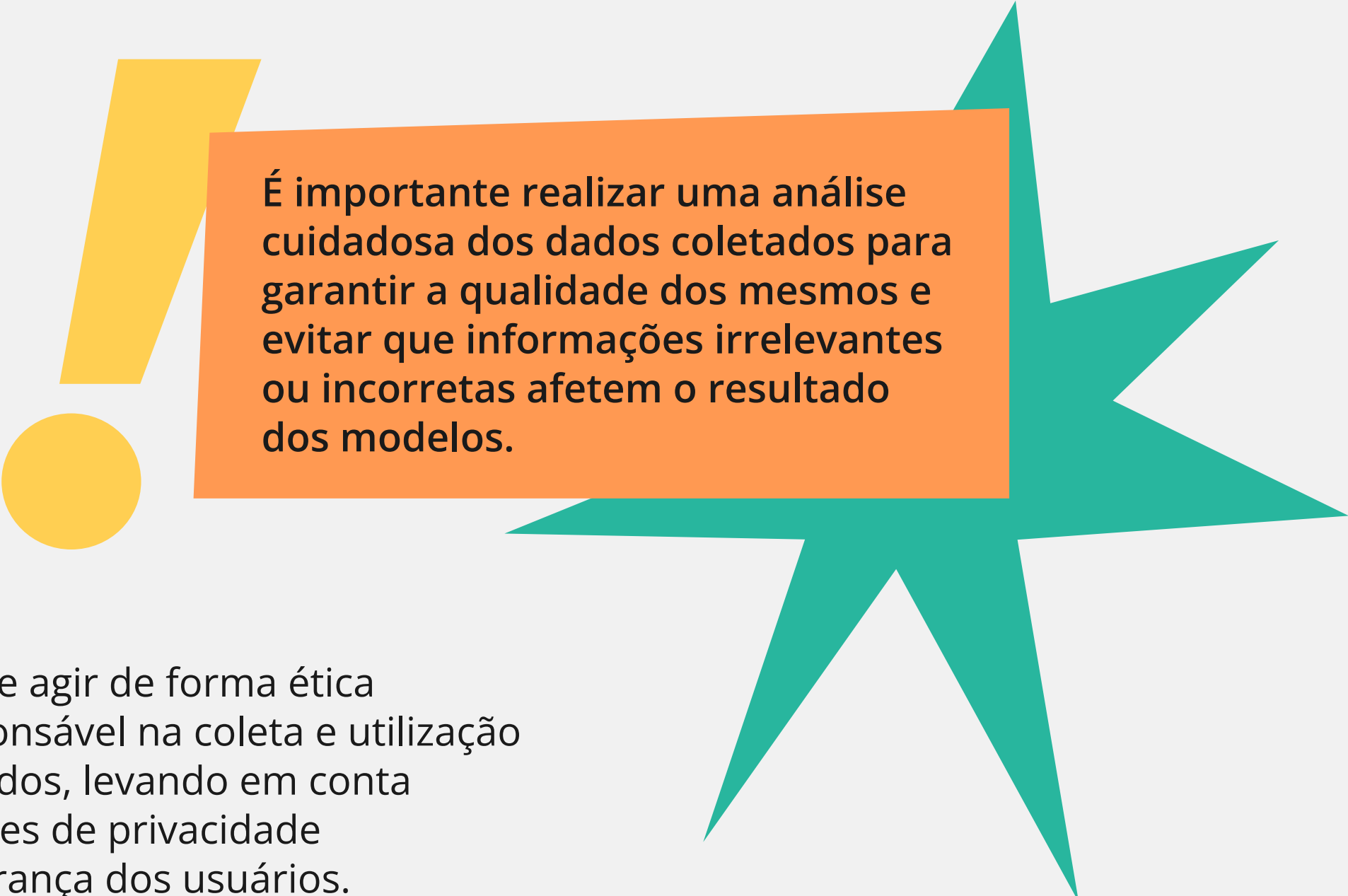
A coleta de dados é uma etapa crucial para ML, pois é a partir dos dados que os algoritmos são treinados para fazer previsões e tomar decisões. Esses dados podem ser obtidos de diversas fontes, como:

- Bancos de dados internos da empresa;
- Dados públicos disponíveis na internet;
- Sensores ou dispositivos IoT (*Internet of Things*).

É importante que os dados coletados sejam relevantes para o problema em questão e que sejam de qualidade, ou seja, **que não contenham erros ou informações duplicadas**.

Coletar dados apresenta alguns desafios e riscos específicos que precisam ser considerados. Entre esses desafios, destacam-se:

- Necessidade de selecionar um conjunto de dados representativo e equilibrado;
- Privacidade dos dados dos usuários;
- Limpeza e pré-processamento dos dados antes de utilizá-los para análise e treinamento de modelos.



É importante realizar uma análise cuidadosa dos dados coletados para garantir a qualidade dos mesmos e evitar que informações irrelevantes ou incorretas afetem o resultado dos modelos.

Deve-se agir de forma ética e responsável na coleta e utilização dos dados, levando em conta questões de privacidade e segurança dos usuários.

3.6 Informações dos dados

Podemos obter várias informações, muitas vezes chamadas de *insights*, analisando conjuntos de dados. É através destes *insights* que encontramos tendências, padrões e correlações no nosso conjunto. Além disso, podemos usar técnicas de análise de dados para extrair diversas informações mais complexas, como por exemplo:

- Agrupamentos (*clusters*);
- Previsões (*forecasting*) de valores futuros;
- Modelagem de relações causais.

Por fim, as informações obtidas a partir dos dados analisados podem ser utilizadas para a tomada de decisão em diferentes áreas. Por exemplo, podemos obter *insights* valiosos sobre o comportamento do usuário, preferências do cliente e eficácia de campanhas de *marketing*, entre outras coisas.

3.7 Pré-processamento dos dados

O pré-processamento de dados consiste em uma série de técnicas que são aplicadas aos dados antes que eles possam ser utilizados para a análise estatística e desenvolvimento de soluções de problemas. O objetivo do pré-processamento também é **melhorar a qualidade dos dados e torná-los mais adequados para serem utilizados pelos modelos**.

O Pandas é uma das bibliotecas mais populares do *Python* para pré-processamento de dados. Ele oferece estruturas de dados flexíveis para trabalhar com tabelas, permitindo a manipulação e limpeza de dados, seleção e filtragem de colunas e linhas, além de operações de agregação e transformação.

Ela também possui recursos para trabalhar com valores ausentes, combinar diferentes conjuntos de dados e lidar com diferentes tipos de dados. Sua interface de usuário intuitiva facilita a visualização e exploração dos dados, bem como a exportação de dados limpos e preparados para análise.

Por fim, o pré-processamento de dados é uma etapa crucial na preparação dos dados para serem utilizados em projetos de Aprendizado de Máquina. Ele permite que os dados sejam limpos, transformados e selecionados para melhorar a qualidade e a adequação dos mesmos aos modelos de Aprendizado de Máquina.

3.7.1 Etapas do pré-processamento de dados

O pré-processamento de dados pode incluir várias etapas, como por exemplo:

- **Limpeza de dados:** envolve a remoção de dados duplicados, a correção de erros e a exclusão de informações irrelevantes;
- **Transformação de dados:** pode incluir a normalização dos dados, para que eles estejam em uma escala semelhante;
- **Codificação de dados categóricos:** transforma-os em números para serem utilizados em modelos matemáticos;
- **Seleção de recursos:** onde apenas os recursos relevantes para o problema são mantidos na amostra;
- **Redução de dimensionalidade:** é outra técnica importante de pré-processamento de dados, especialmente em conjuntos de dados com muitas variáveis. A ideia é reduzir a dimensão do conjunto de dados, mantendo as informações mais relevantes;
- **Análise de componentes principais:** do inglês, PCA, é uma das técnicas mais comuns que projeta os dados em um espaço de menor dimensão, mantendo as características mais importantes.

Aula 4 - Ética e sensibilidade dos dados



Foto de Onur Buz,
disponível na Unsplash.
Adaptada pelo autor.

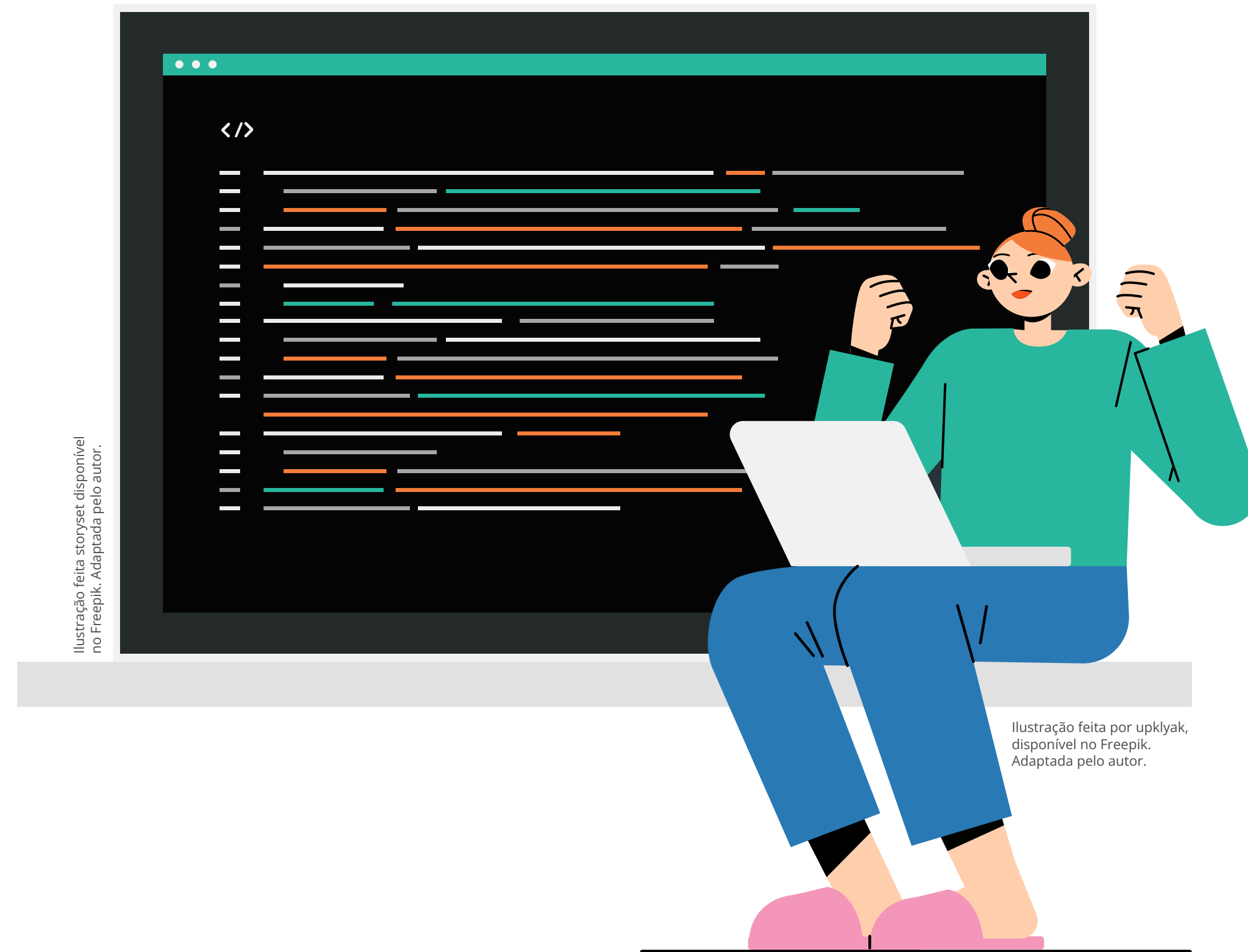
Foto de Marissa Lewis, disponível
na Unsplash. Adaptada pelo autor.

A ética e a sensibilidade de dados referem-se à responsabilidade dos profissionais de dados em lidar com informações que possam afetar a privacidade, a segurança ou a dignidade de pessoas, grupos ou organizações.

A ética de dados envolve **o tratamento justo e transparente das informações, incluindo a coleta, armazenamento, processamento, análise e compartilhamento de dados**. Isso inclui a necessidade de obter consentimento informado, proteger a privacidade e a segurança dos dados, e evitar discriminação ou viés em relação às informações coletadas.

A sensibilidade de dados é a consideração das implicações sociais e políticas que os dados podem ter. Ela aborda questões como o uso adequado dos dados para evitar danos e preservar os direitos e a privacidade das pessoas e comunidades envolvidas. Além disso, a sensibilidade de dados também envolve garantir que os dados coletados representem de forma justa e precisa as populações e comunidades, levando em conta questões de inclusão e diversidade.

Esses são aspectos importantes para garantir que o uso de dados seja benéfico para a sociedade e respeite os valores éticos.





**TIC em
trilhas**

Foto de Chase Clark, disponível na
Unsplash. Adaptada pelo autor.

Foto disponível no
FreePik. Adaptada
pelo autor.

Aula 5 - Visualização de dados

A visualização de dados é uma técnica utilizada para representar informações em gráficos, tabelas e outros formatos visuais que tornam mais fácil entender e analisar grandes quantidades de dados. A ideia é que, ao invés de olhar para grandes tabelas com dados, os usuários possam interpretar as informações de forma **mais rápida e intuitiva** através de gráficos e outras representações visuais.

As visualizações de dados podem ser criadas a partir de diversas fontes de dados, incluindo planilhas, bancos de dados e outras fontes de informação. Elas podem ser utilizadas para diversos propósitos, como identificar padrões em grandes quantidades de dados, entender a distribuição dos dados e detectar tendências ao longo do tempo.



Existem muitas ferramentas disponíveis para criar visualizações de dados, desde planilhas e programas de gráficos simples, até *softwares* avançados de análise de dados. O objetivo final é que as visualizações de dados permitam que os usuários entendam e comuniquem informações complexas, tornando o processo de análise de dados mais eficiente e acessível.

Ao criar visualizações de dados, é importante considerar a relação entre:

- **Expressividade:** refere-se à capacidade da visualização de transmitir informações de maneira clara;
- **Efetividade:** refere-se à capacidade da visualização de comunicar informações precisas e úteis.

Portanto, é importante escolher visualizações que equilibrem esses dois fatores.

5.1 Visualizações fundamentais

5.1.1 Gráfico de dispersão

O gráfico de dispersão é uma das visualizações mais comuns e simples. Ele mostra a **relação entre duas variáveis numéricas**, onde cada ponto no gráfico representa uma observação com valores para ambas as variáveis, representadas nos eixos X e Y. Apesar de sua simplicidade, o padrão de dispersão dos pontos no gráfico pode mostrar uma relação positiva ou negativa, além de ser útil para identificar pontos incomuns e discrepantes.

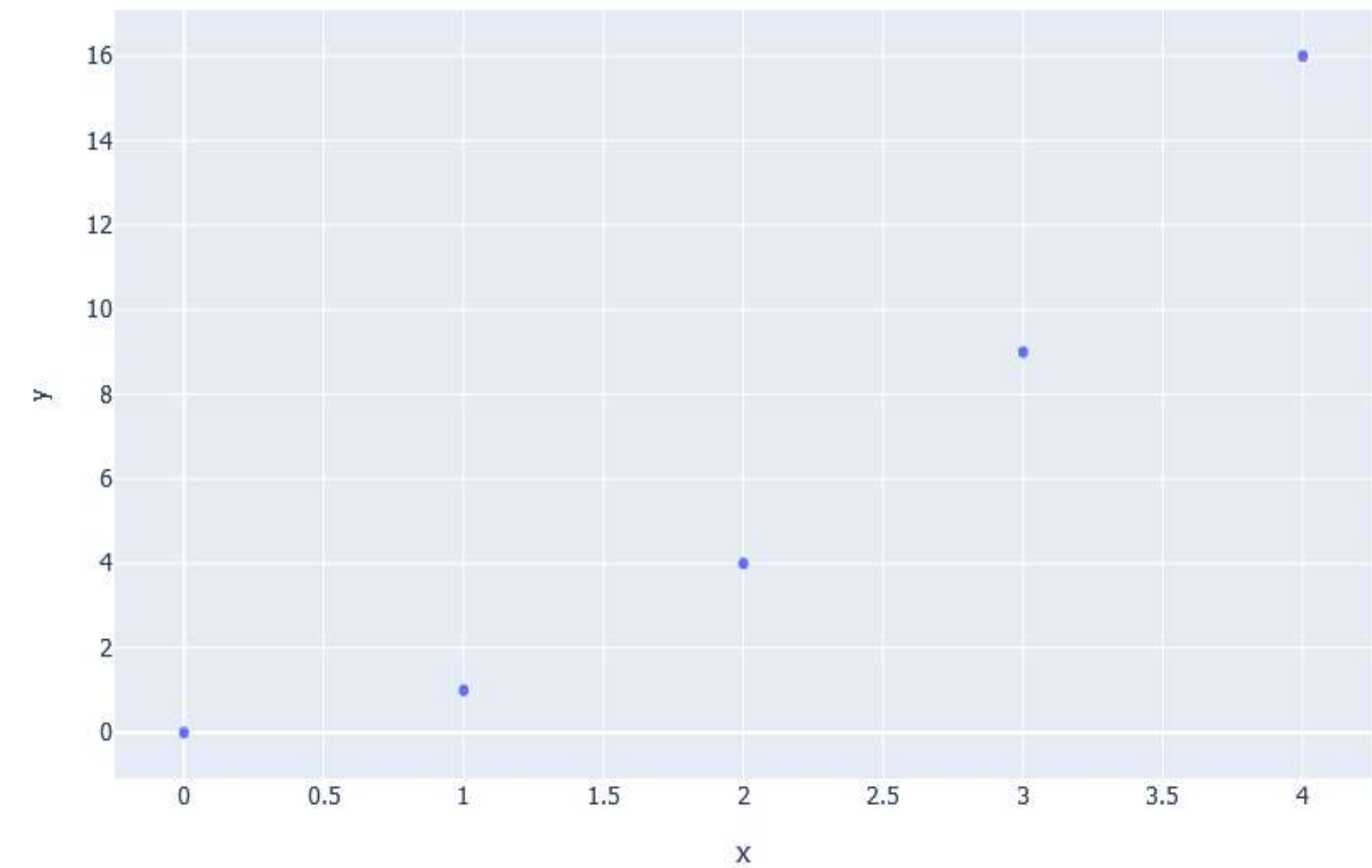


Figura 3 – Visualização em gráfico de dispersão. Fonte: feito pelo autor.

5.1.2 Gráfico de barras

O gráfico de barras é uma representação visual de dados categóricos em que as categorias são plotadas no eixo X e as contagens ou frequências são plotadas no eixo Y. Esse tipo de gráfico é útil para **comparar a distribuição de uma variável em diferentes categorias ou para mostrar a evolução de uma variável ao longo do tempo**. No exemplo abaixo, os eixos X e Y representam o número de anos e o tamanho da população do Canadá, em milhões.

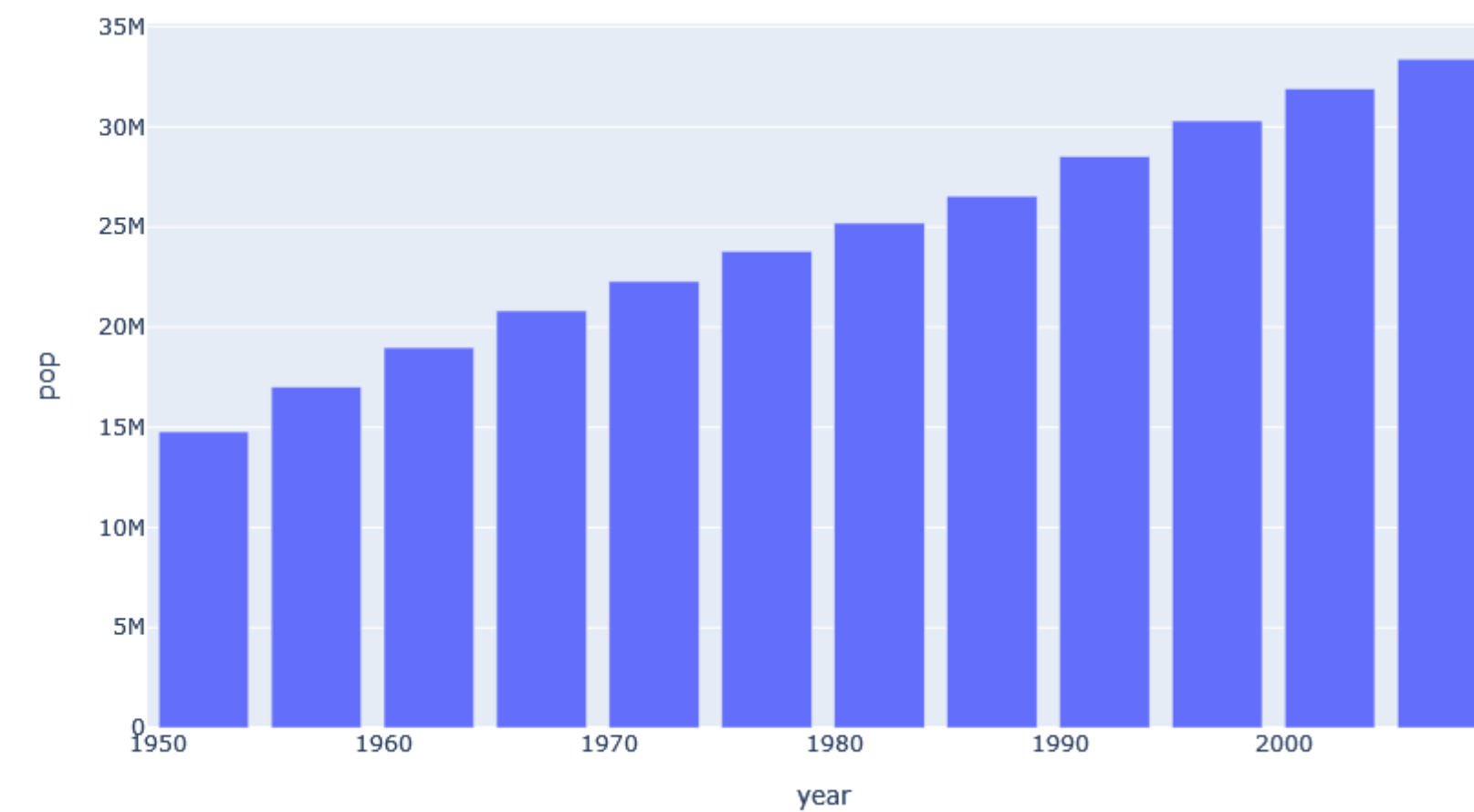


Figura 4 – Visualização em gráfico de barras. Fonte: feito pelo autor.

5.1.3 Gráfico de linha

O gráfico de linhas é criado traçando pontos para cada valor de dados em um eixo X e Y, e então conectando os pontos em uma linha suave. Assim, é frequentemente usado para **mostrar tendências, padrões ou mudanças ao longo do tempo**, permitindo que os usuários identifiquem facilmente o aumento ou diminuição de valores em uma série de dados. No exemplo a seguir, podemos comparar a expectativa de vida da população da Austrália (azul) e da Nova Zelândia, no passar dos anos.

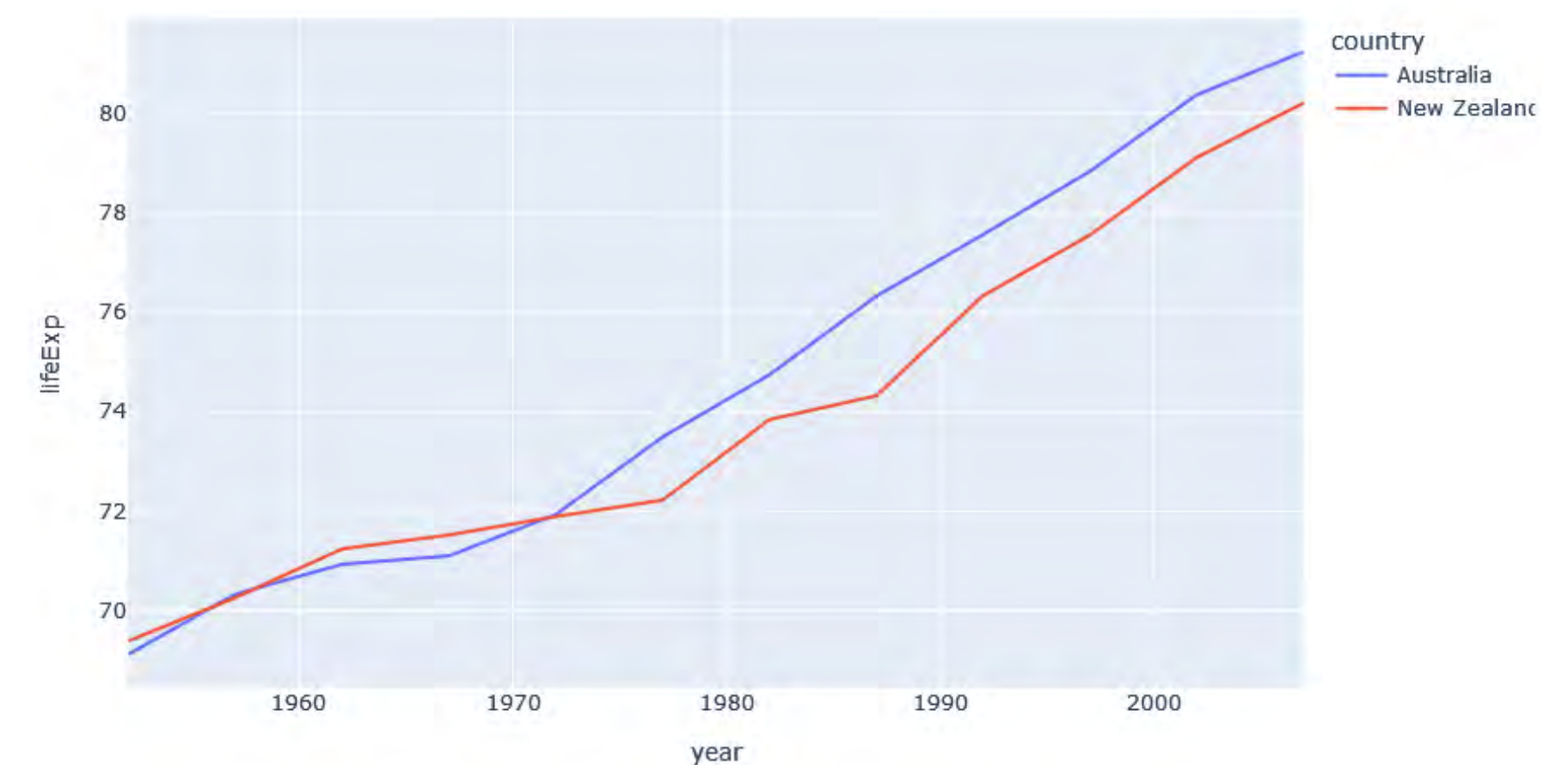


Figura 5 – Visualização em gráfico de linhas. Fonte: feito pelo autor.

5.1.4 Gráfico de pizza

Um gráfico de pizza é um tipo de gráfico circular que é utilizado para **representar a proporção de cada categoria em um conjunto de dados**. O gráfico é dividido em fatias, onde cada fatia representa uma categoria e o tamanho da fatia é proporcional à quantidade de dados naquela categoria. É uma forma simples e eficaz de visualizar dados categóricos. No gráfico abaixo podemos ver a porcentagem da população de cada país americano em relação à população total do continente.

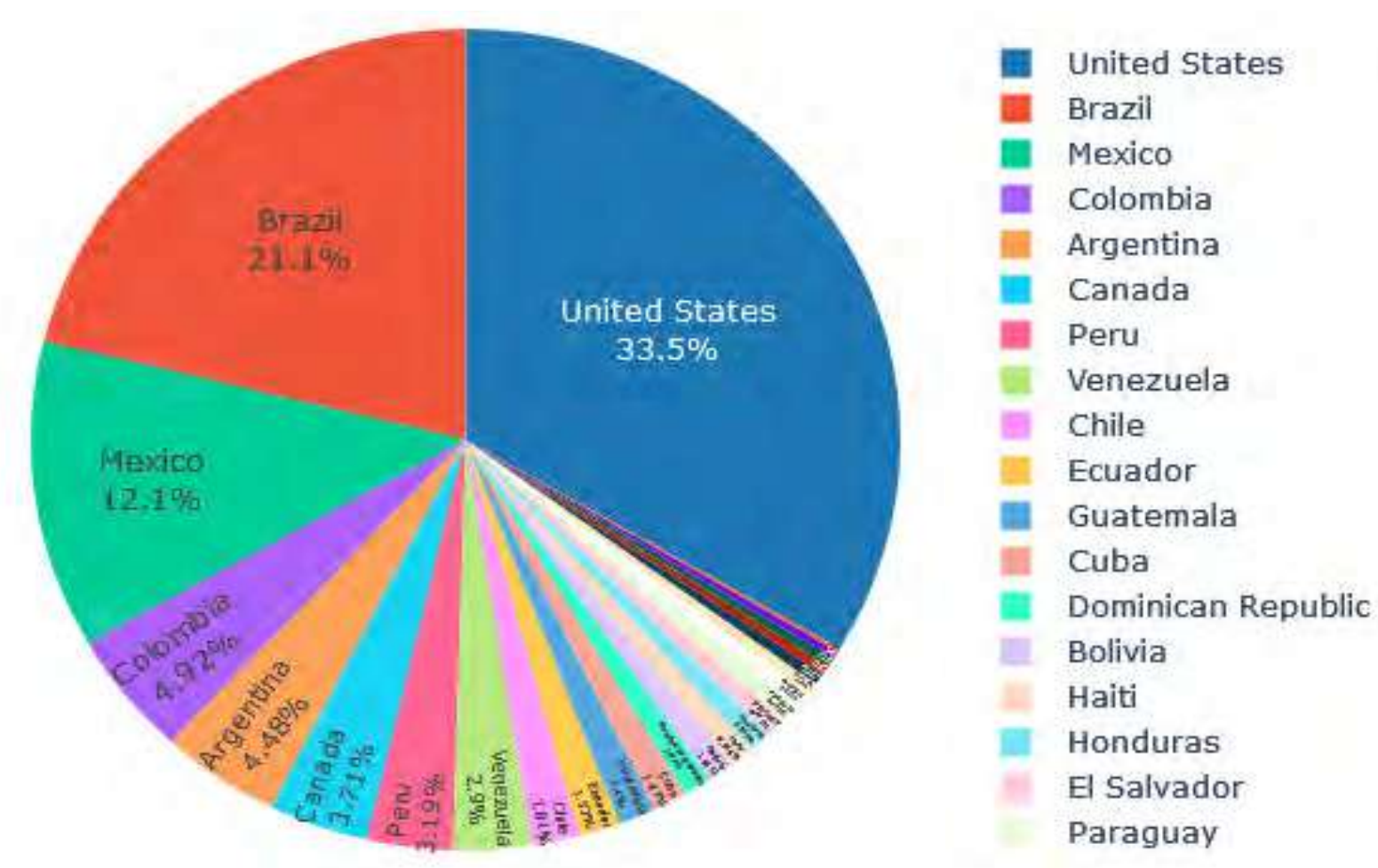


Figura 6 – Visualização em gráfico de pizza. Fonte: feito pelo autor.

5.1.5 Histograma

Um histograma é um gráfico de barras que representa **a distribuição da frequência de uma variável contínua**. No eixo X, os valores são divididos em intervalos de igual tamanho, chamados de *bins*, e no eixo Y, é mostrada a frequência de ocorrência dos valores em cada *bin*. O histograma é uma ferramenta útil para visualizar a forma da distribuição de um conjunto de dados, como a presença de picos, assimetrias e a dispersão dos valores.

Embora os gráficos de barras e histogramas tenham visualizações semelhantes, eles diferem na forma como os dados são representados. Um gráfico de barras é usado para mostrar a contagem ou frequência de diferentes categorias ou grupos discretos de dados. Cada barra representa uma categoria e a altura da barra indica a frequência ou contagem da categoria correspondente. Por outro lado, um histograma é usado para mostrar a distribuição de uma variável contínua. Os dados são divididos em intervalos e as alturas das barras indicam a frequência ou densidade de dados dentro de cada intervalo. Em resumo, enquanto os gráficos de barras são usados para dados categóricos, o histograma é usado para dados contínuos.

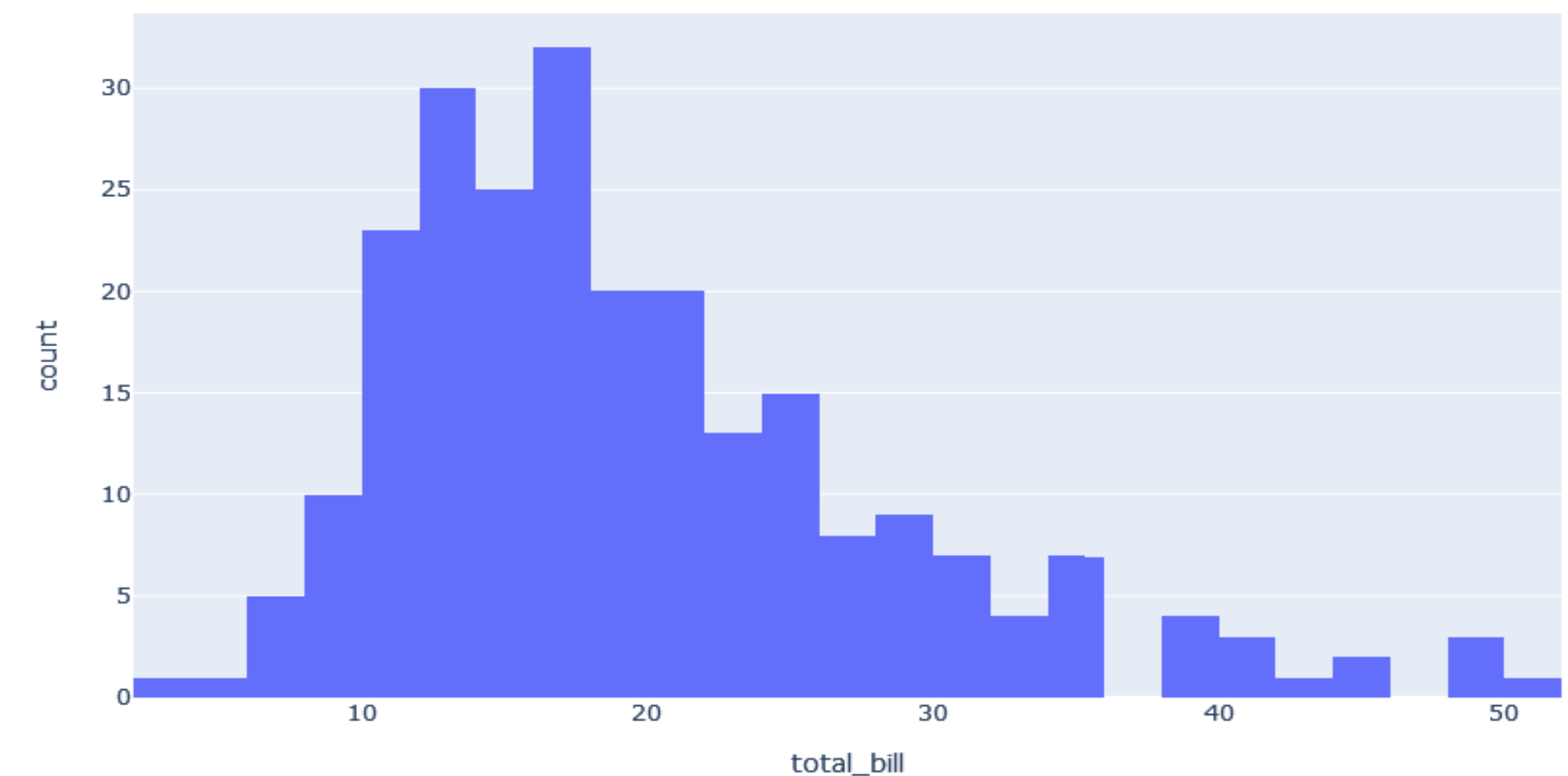


Figura 7 – Visualização em histograma. Fonte: feito pelo autor.

5.1.6 Gráfico de calor

Um gráfico de calor é uma representação visual de **dados que usa cores para mostrar a intensidade de uma valor em uma matriz ou tabela**. Cada célula da matriz é colorida com uma cor diferente, dependendo do valor que ela contém. No exemplo abaixo, vemos um gráfico de 3x3 que mostra a quantidade de medalhas, ouro, prata e bronze, por país (eixo Y).

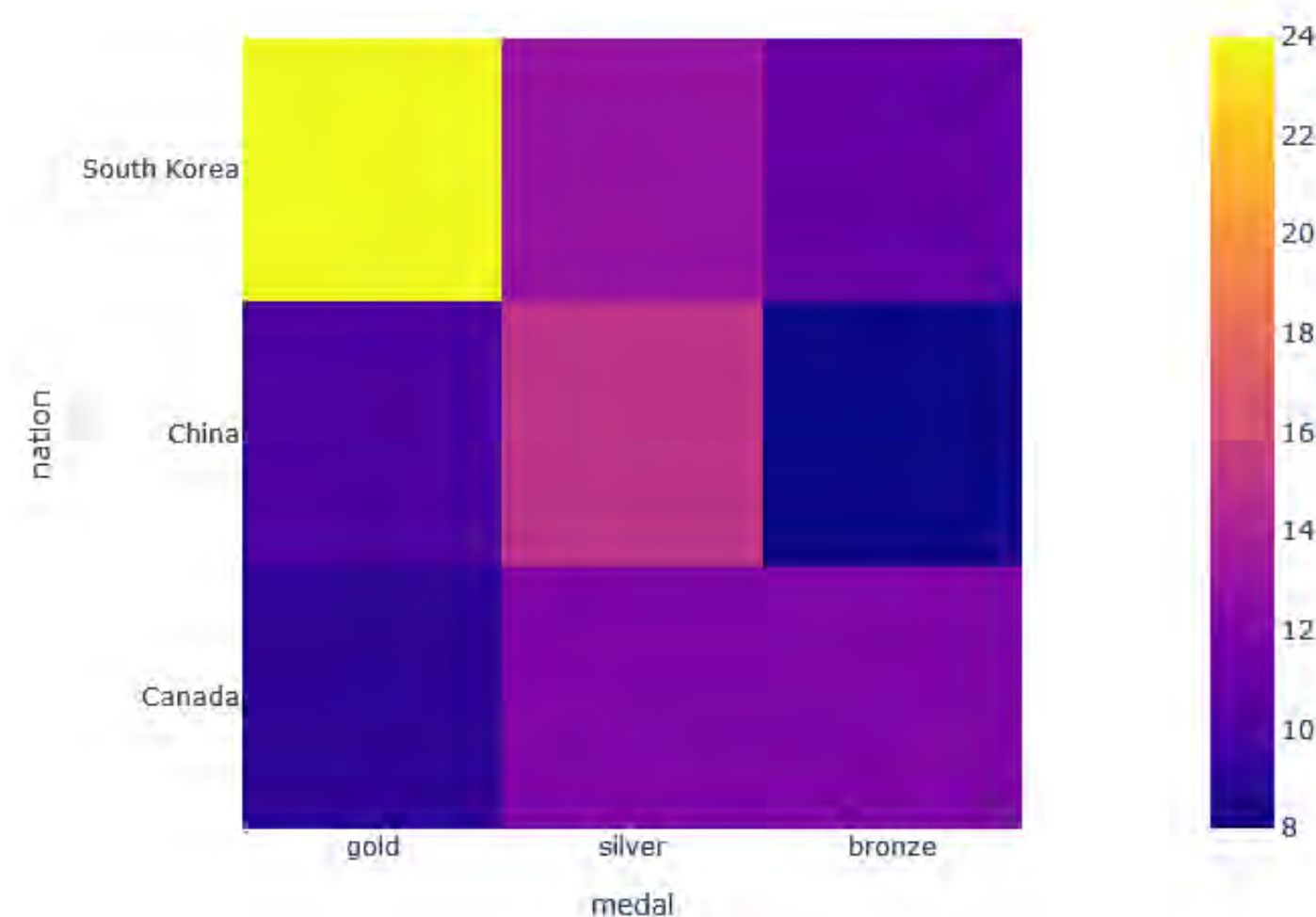


Figura 8 – Visualização em gráfico de calor. Fonte: feito pelo autor.

5.1.7 Bibliotecas de visualização

A visualização de dados em *Python* é facilitada por várias bibliotecas especializadas que devem ser utilizadas de acordo com as **necessidades do projeto**. A escolha da biblioteca dependerá do tipo de visualização que se deseja criar e do nível de interatividade que se pretende atingir. Veja a seguir algumas opções de bibliotecas:

- **Matplotlib:** é uma das bibliotecas mais populares para visualização de dados da linguagem, sendo utilizada para criar gráficos de linhas, barras, dispersão, histogramas e muito mais. Ela é altamente personalizável e permite que os usuários controlem praticamente todos os aspectos visuais de seus gráficos;
- **Seaborn:** é outra biblioteca popular de visualização de dados, que se concentra em gráficos estatísticos e visuais complexos, como mapas de calor (*heatmaps*), gráficos de dispersão e de densidade. Ele é construído “em cima” do *Matplotlib* e fornece uma API mais simples e fácil de usar para visualização de dados estatísticos;
- **Plotly:** é uma biblioteca de visualização mais avançada que permite criar gráficos interativos e dinâmicos. Ela suporta vários tipos de gráficos, como gráficos de barras, linhas, pizza e muitos outros. O *Plotly* é adequado para projetos que exigem interatividade, como painéis de dados em tempo real e dashboards personalizados.

Explore mais!

O filme "Jogo da Imitação" conta a história de Alan Turing precursor dos estudos sobre inteligência artificial e sua contribuição para a invenção do computador.

O filme "Estrelas Além do Tempo" mostra a história de três matemáticas que se tornam pioneiras na programação.

"A Saga dos Computadores", do canal Manual do Mundo, é uma série de vídeos que relata a origem e funcionamento dos computadores. Assista clicando [aqui.](#)



Referências Bibliográficas

SCIKIT LEARN. 1.1. Linear Models — scikit-learn 0.24.0 documentation. Disponível em: https://scikit-learn.org/stable/modules/linear_model.html. Acesso em: abril de 2023.

SCIKIT LEARN. 1.4. Support Vector Machines — scikit-learn 0.20.3 documentation. Disponível em: <https://scikit-learn.org/stable/modules/svm.html>. Acesso em: abril de 2023.

AURÉLIEN GÉRON. Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems. 2. ed. [s.l.] O'Reilly Media, Inc., 2019. Acesso em: abril de 2023.

PLOTLY. Plotly Python Graphing Library. Disponível em: <https://plotly.com/python/>. Acesso em: abril de 2023.

MARQUES, R. F. Aprendizado de máquina: subáreas e aplicações - Aquarela Analytics. Disponível em: <https://www.aquare.la/aprendizado-de-maquina-subareas-e-aplicacoes/>. Acesso em: abril de 2023.

pandas documentation — pandas 1.0.1 documentation. Disponível em: <https://pandas.pydata.org/docs/>. Acesso em: abril de 2023.

Gallery — Matplotlib 3.4.2 documentation. Disponível em: <https://matplotlib.org/stable/gallery/index.html>. Acesso em: abril de 2023.

GOPINATH REBALA; AJAY RAVI; SANJAY CHURIWALA. An Introduction to Machine Learning. [s.l.] Cham Springer International Publishing Imprint, Springer, 2019. Acesso em: abril de 2023.