Análise de dados

MANUAL DO PROFISSIONAL DE DADOS

Fernanda Santos

SOBRE A AUTORA

Olá, meu nome é Fernanda, sou formada em Gestão da Informação e apaixonada por esse mundo da análise de dados. Comecei no *Excel*, depois entrei no mundo do *Business Intelligence*, utilizando o ***Power BI***, e atualmente sou Cientista de dados.Este *e-book* tem o objetivo de abordar alguns conceitos nessa área, logo espero ajudar você, de alguma forma, com o conteúdo que preparei.

INTRODUÇÃO

Na atualidade, é evidente, que nada poderia funcionar sem uma quantidade significativa de informação, como elemento que impulsiona os fenômenos sociais e que é por eles impulsionada. Uma gestão da informação eficiente gera inteligência competitiva, tanto em organizações privadas, como em públicas. Logo, a ação de transformar dados em informações significativas melhora a tomada de decisão e consequentemente coloca a empresa em um patamar competitivo diferenciado. Este manual abordará alguns conceitos na área de análise de dados e business intelligence para ajudar a todos que gostariam de conhecer um pouco mais sobre esse mundo tão imenso.

Dado Vs. Informação

Para um melhor entendimento, vamos relembrar o que é dado e o que é informação?

Dado é a menor unidade de informação, como por exemplo, bola. Ao falar apenas a palavra bola para alguém, a pessoa certamente não entenderá a mensagem que você está tentando passar, pois, bola é apenas um dado, e, precisamos dar sentido a mensagem que estamos tentando passar. Informação, é quando aplicamos o dado bola, que vimos anteriormente, em um contexto. Sendo assim, se agora ao invés de falar apenas a palavra bola, a gente fala que João jogou bola hoje, acabamos de transformar os seguintes dados: "João", "jogou", "bola", "hoje", em informação, e agora sabemos que João jogou bola hoje. Por fim, podemos dizer que informação é quando estruturamos os nossos dados, fazendo com que eles façam sentido.

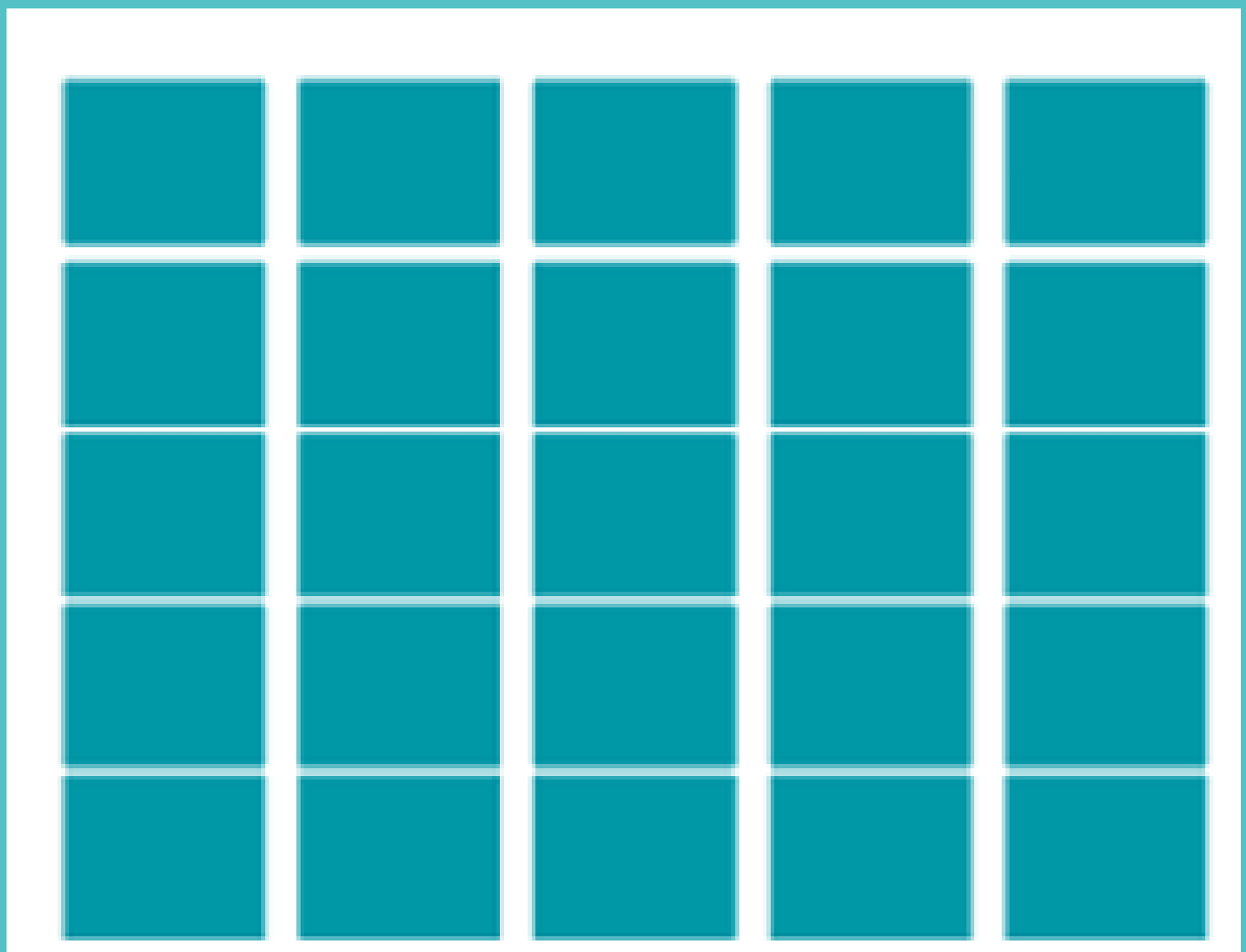
Dados Estruturados x Dados não estruturados

Dados estruturados, como o nome já diz, são dados que seguem, possuem uma estrutura rígida e seguem um padrão, por exemplo, planilhas, banco de dados SQL, arquivos csv, etc...

Com a chegada do big data, ou seja, o grande volume de dados, passamos a consumir dados de várias outras fontes, dos mais variados formatos, como dados de redes sociais(Facebook, Instagram, Twitter, etc...) Emails, Imagens, arquivos de áudio, entre outros, esses dados são o que podemos chamar de dados não estruturados, e, como podemos imaginar não possuem uma estrutura mínima nem seguem um padrão.

Nas figuras abaixo temos uma representação de dados estruturados e não estruturados.

Dados Estruturados



Dados não estruturados

O que é o Business Intelligence?

Business Intelligence ou simplesmente BI,é o processo de coletar, organizar e analisar dados para auxiliar na tomada de decisão. Assim, o BI não é uma ferramenta, o BI é um conceito, porém, existem ferramentas de BI no mercado, amplamente utilizadas para auxiliar em todas as etapas da análise de dados. Abaixo as 3 pricipais ferramentas nessa área eleitas pela Gartner,uma consultoria que atua no mercado de TI desde 1970.

1) Tableau
2) Power BI
3) Qlik

E o BIG DATA, o que é?

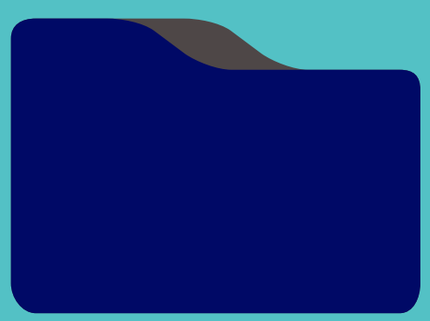
Big Data ou Grande volume de dados, compreende, como o próprio nome diz, a análise de um grande volume de dados. Não é difícil entender o cenário em que o conceito se aplica: trocamos milhões de e-mails por dia; milhares de transações bancárias acontecem no mundo a cada segundo; soluções sofisticadas gerenciam a cadeia de suprimentos de várias fábricas neste exato momento; operadoras registram a todo instante chamadas e tráfego de dados do crescente número de linhas móveis no mundo todo. Se uma empresa souber como utilizar os dados que tem em mãos, poderá entender como melhorar um produto, como criar uma estratégia de marketing mais eficiente, como cortar gastos, como produzir mais em menos tempo, como evitar o desperdício de recursos, como superar um concorrente, como disponibilizar serviços para a um cliente especial de maneira satisfatória e assim por diante.

Os 4 Principais tipo de Análise de Dados

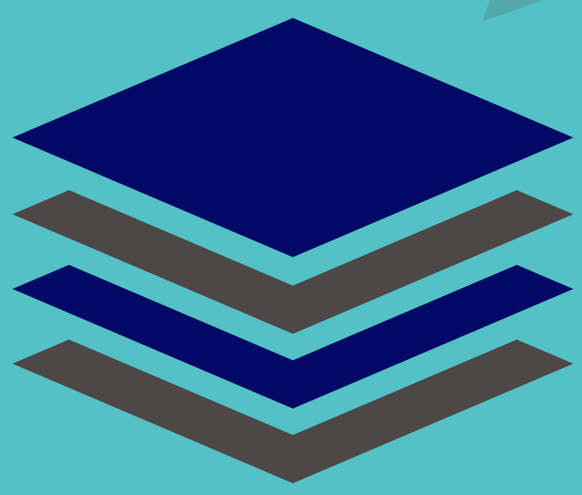
Análise Descritiva	O que Aconteceu?
Análise Diagnóstica	Por que Aconteceu?
Análise Preditiva	O que Acontecerá?
Análise Prescritiva	Como fazer com que Aconteça?

O *Business intelligence* se encaixa nas 2 primeiras análises enquanto o Big Data engloba as 4 análises.

Etapas Da Análise de Dados



Tratamento



Modelagem

Visualização

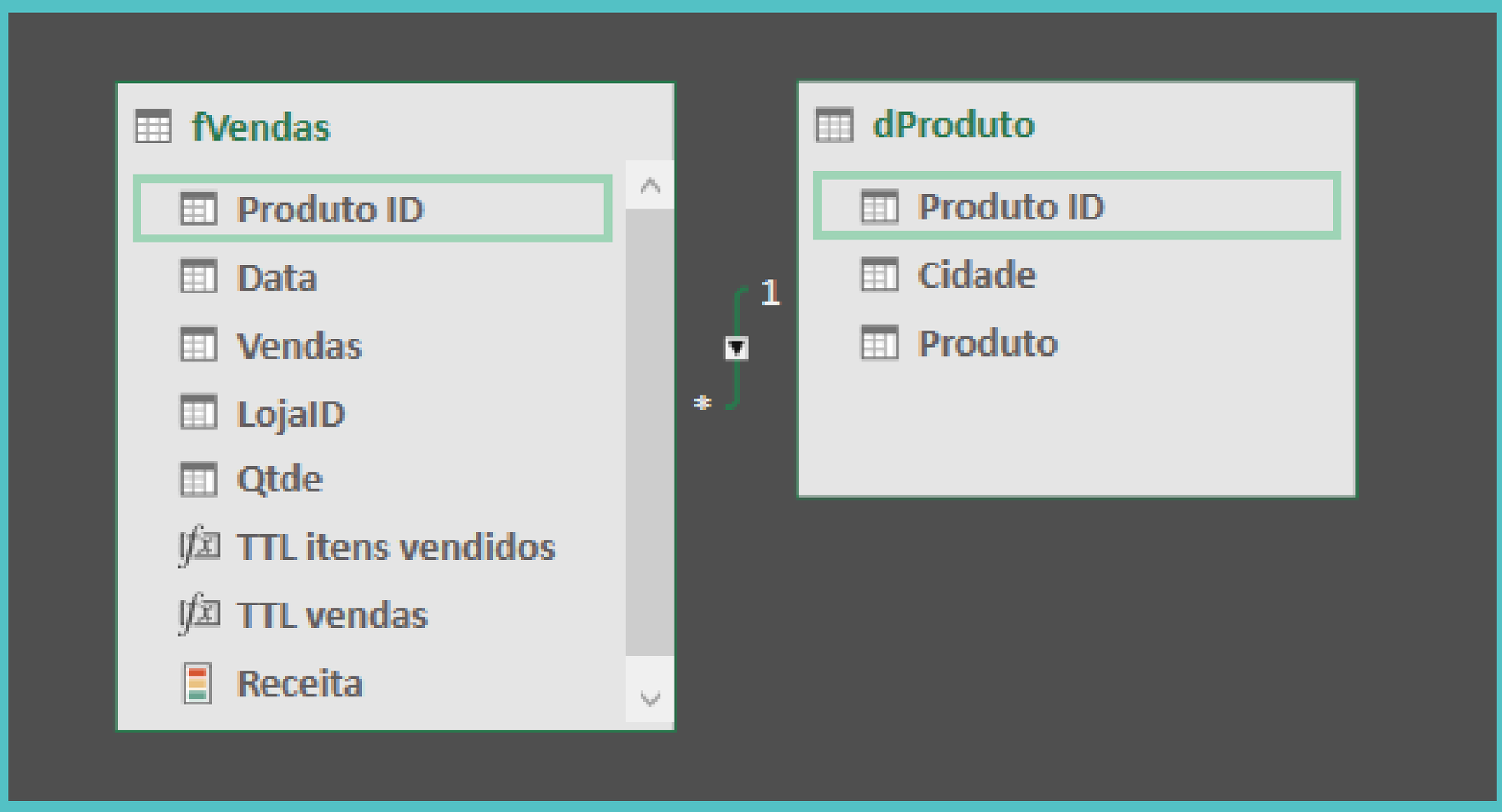
Tratamento

É nessa etapa que é aplicado o *ETL*(*Extract,Transform, Load*),ou seja, a extração, transformação e carregamento dos dados, nessa etapa é definido quais serão as tabelas dimensões nas quais iremos enxergar os fatos. Calma, você não sabe o que é uma tabela fato e uma tabela dimensão? Eu explico:
Tabela Fato: Tabela onde contém valores, exemplo: Tabela de vendas, Tabela com Quantidade de Reclamações, Quantidade de Ocorrência etc...

Tabela Dimensão:Tabela onde contém as dimensões que iremos enxergar o fato, exemplo:
Dimensão tempo: vou querer mostrar minhas vendas por Ano/Mês.
Dimensão Produto:Vou querer mostrar minhas vendas por produto, etc...

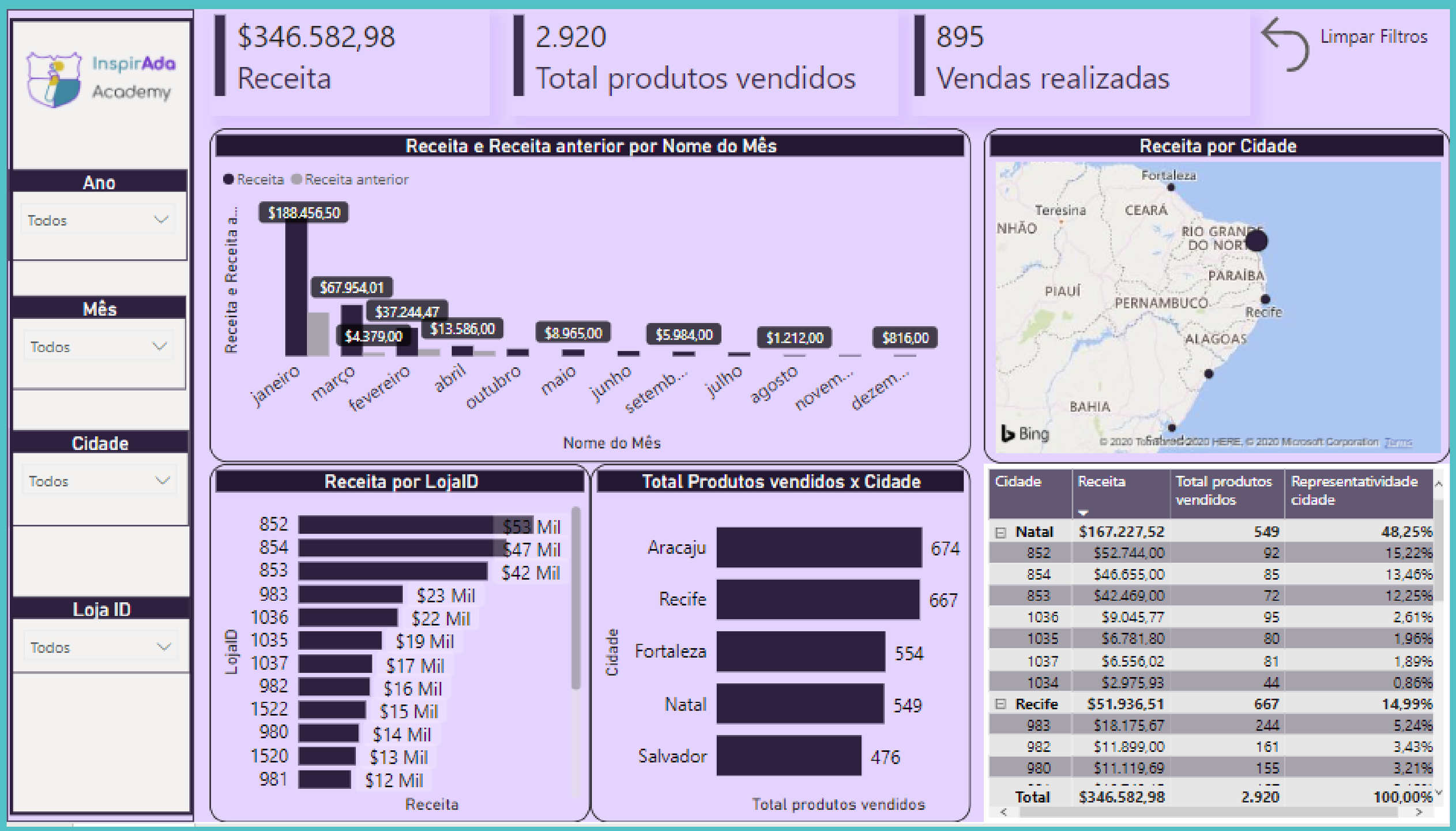
Modelagem

Agora, que você já sabe o que é uma tabela fato e uma tabela dimensão, vamos a etapa da modelagem, aqui é criado o relacionamento entre as tabelas para que possamos obter êxito na elaboração dos painéis, lembra da tabela de vendas e da tabela de produto? é nessa etapa que criamos a relação entre o que as 2 tabelas possuem em comum, por exemplo: se na tabela de vendas eu tenho o ID de cada produto vendido e na tabela de produtos eu também tenho esse ID,então eu posso relacioná-las para obter uma maior performance no modelo de dados.
Segue Exemplo abaixo:



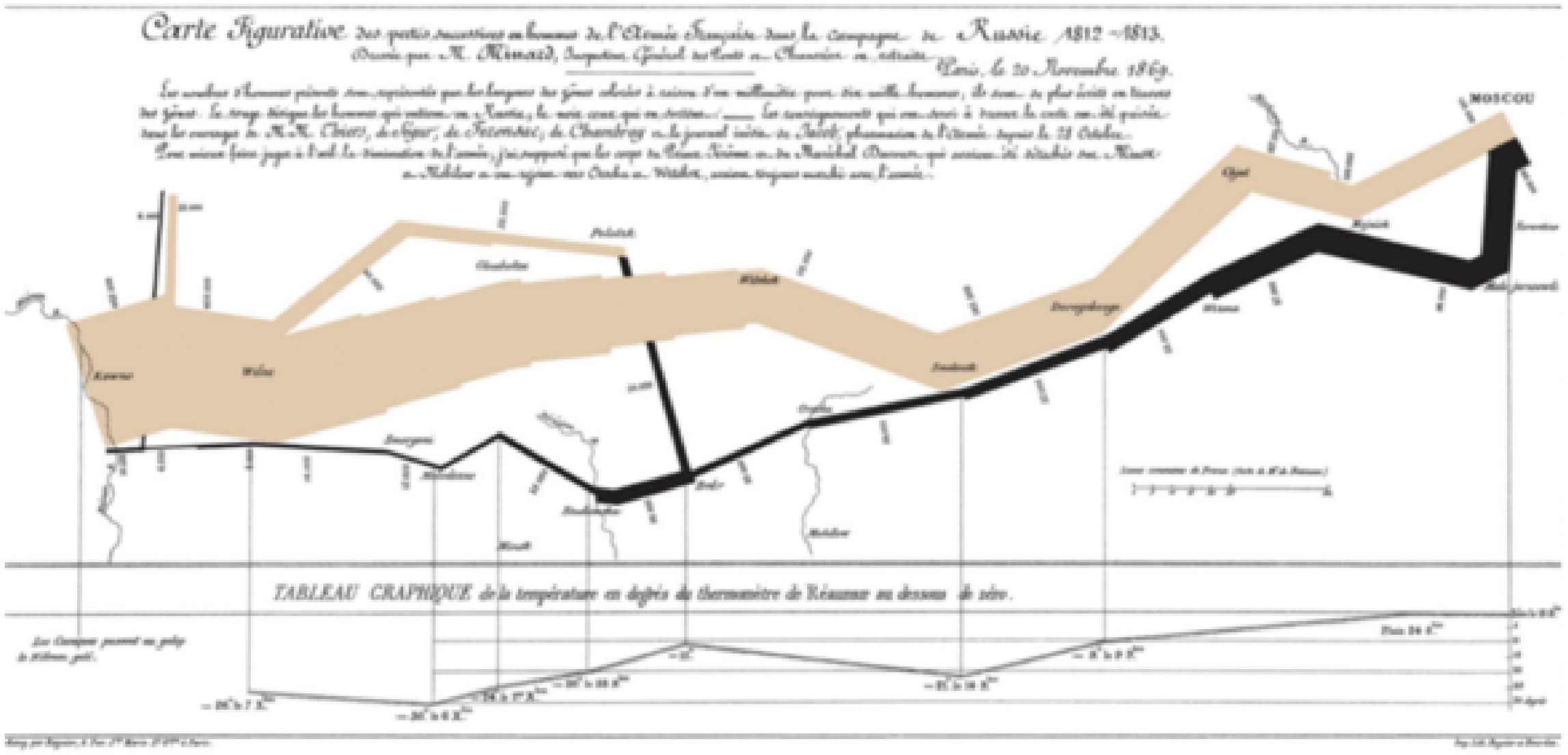
Visualização

Esta é a última etapa da análise e dados e a mais rápida, depois de tratar e modelar nossos dados estamos prontos para criar nossas visualizações, aí é só criar nossos gráficos e tabelas que pretendemos mostrar para o usuário final, a atenção aqui é em como mostrar os dados para que o usuário não tenha dúvida e não faça uma interpretação errada da informação ali mostrada. Segue abaixo um exemplo de um painel de Vendas criado no Power BI.



Com a visualização de dados é possível captar conceitos e padrões, se você deseja se tornar um cientista de dados, conseguir apresentar suas ideias de forma clara e concisa será uma habilidade muito importante. A visualização de dados não é uma área recente, existe há séculos, desde os mapas e diagramas do século XVII até a invenção do gráfico de pizza, por volta dos anos 1800. E, ainda que os desenhos de antigamente sejam menos atrativos do que os atuais, mesmo assim, algumas das primeiras visualizações foram bem surpreendentes e poderosas. Uma das visualizações de dados históricos mais famosas foi construída por Charles Joseph Minard, ele era um engenheiro civil francês e famoso por representar dados numéricos em mapas geográficos. Na figura abaixo Minard descreve de forma gráfica a jornada de Napoleão Bonaparte, enquanto ele marchava em direção à Rússia, para sua campanha russa de 1812.

Joseph Minard 1861



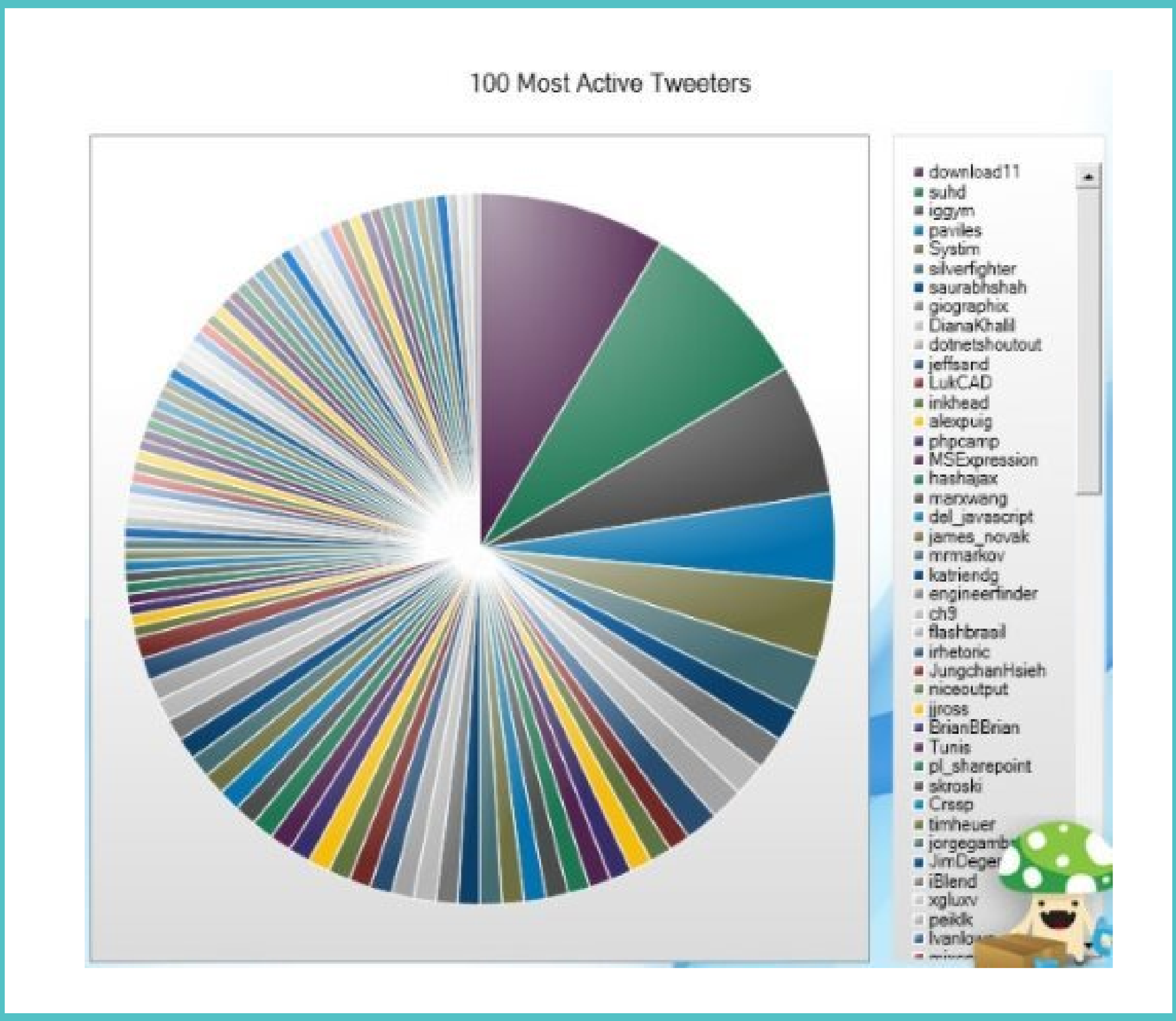
A imagem mostra o exército de Napoleão partindo da fronteira Polonesa-Russa. Uma faixa espessa mostra o tamanho do seu exército em pontos geográficos específicos durante seu avanço e retirada. Também é possível notar na imagem a perda considerável do exército de Napoleão sobre o avanço em Moscou.

O exemplo que acabamos de ver construído por Minard é de uma boa visualização de dados, posteriormente esse tipo de gráfico para ilustração de fluxos passou a ser chamado de Diagrama de Sankey.

Nem sempre vamos nos deparar com boas visualizações, uma boa visualização de dados não se trata apenas de criar o gráfico, uma boa visualização de dados precisa comunicar de forma clara a mensagem que se deseja passar ao receptor. Vamos agora ver alguns exemplos de visualizações ruins.

Podemos obsevar na visualização acima que a barra de 47º às 5AM está mais alta que a barra de 47º das 7AM e, também é mais alta que a de 48º às 9AM. Outro ponto é que a barra de 50º às 5PM está mais alta que a barra de 51º às 3PM.

Essa visualização pode levar o receptor a conclusões erradas com relação a temperatura em determinado horário do dia. Vamos a outro exemplo:



Vocês entenderam o gráfico acima? Se sim, me expliquem por favor. O gráfico de pizza tem o objetivo de mostrar as partes de um todo, no caso acima foram utilizadas muitas categorias, fazendo com que cause uma poluição gráfica, além de que, não é possível extrair nenhum insight sobre o que está sendo mostrado.

Os gráficos utilizados como exemplos de como não fazer visualização de dados foram retirados do site https://viz.wtf. O site tem vários exemplos de visualizações que não fazem sentido. Por fim, para criar uma boa visualização de dados, podemos responder algumas perguntas antes de colocar a mão na massa. Estas perguntas nos guiarão no processo conceitual de visualização de dados, sendo estas:

1 - Que tipos de dados você tem?

2 - O que você pretende estudar sobre os dados?

3 - Qual a melhor maneira de visualizar estes dados?

Para responder a primeira pergunta e te ajudar a escolher qual tipo de gráfico deve ser utilizado, deixo aqui a dica do site https://www.data-to-viz.com que ajuda você a escolher o gráfico mais apropriado para os seus dados.

Data Science

Agora que já conhecemos mais o processo de análise de dados, vamos avançar e falar um pouco sobre Data Science. Data Science ou Ciência de dados, é a ciência que estuda o processo de coleta, tratamento, transformação e análise de dados. Podemos entender a Ciência de dados como uma área multidisciplinar, que envolve Ciência da computação, Matemática e estatística e o domínio do negócio. Sendo assim, uma equipe de Data Science precisa ser composta por pessoas com habilidades de programação, domínio da estatística e pessoas que tenham conhecimento do negócio ao qual está sendo estudado. As linguagens de programação mais utilizadas em Data Science são Python e R, porém, não são as únicas. Você pode estar se perguntando qual a diferença entre Data Science e Business Intelligence, já que os conceitos são bem parecidos. o Business Intelligence analisa os dados históricos, sempre olhando pelo retrovisor, enquanto que Data Science busca realizar análises preditivas para responder o que provavelmente irá acontecer.

E o Machine Learning?

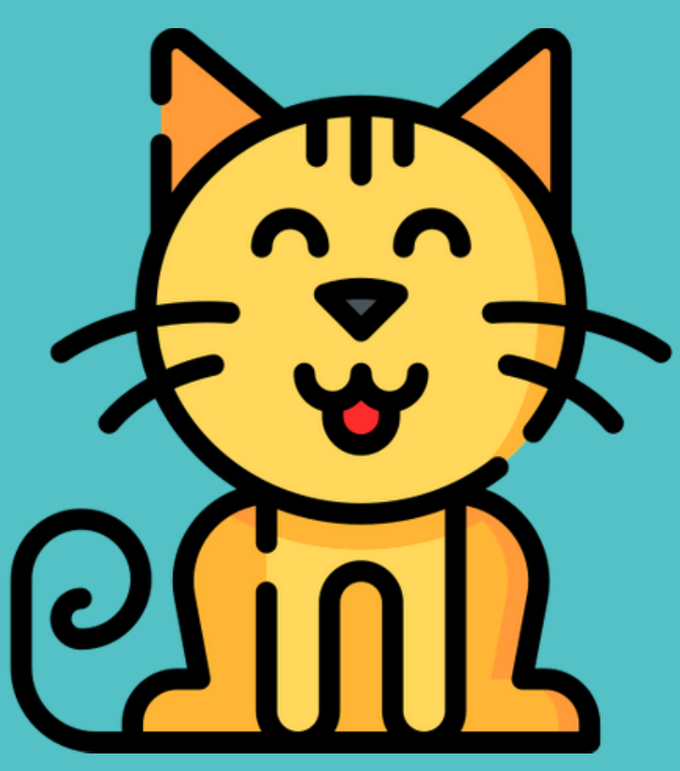
Machine Learning, ou aprendizado de máquina é um subcampo da Ciência da computação e tem como objetivo reconhecer padrões em nossos dados para nos auxiliar com nossas análises preditivas. Os principais tipos de aprendizado de máquina são:
Supervisionado
Não supervisionado
Aprendizado por reforço.

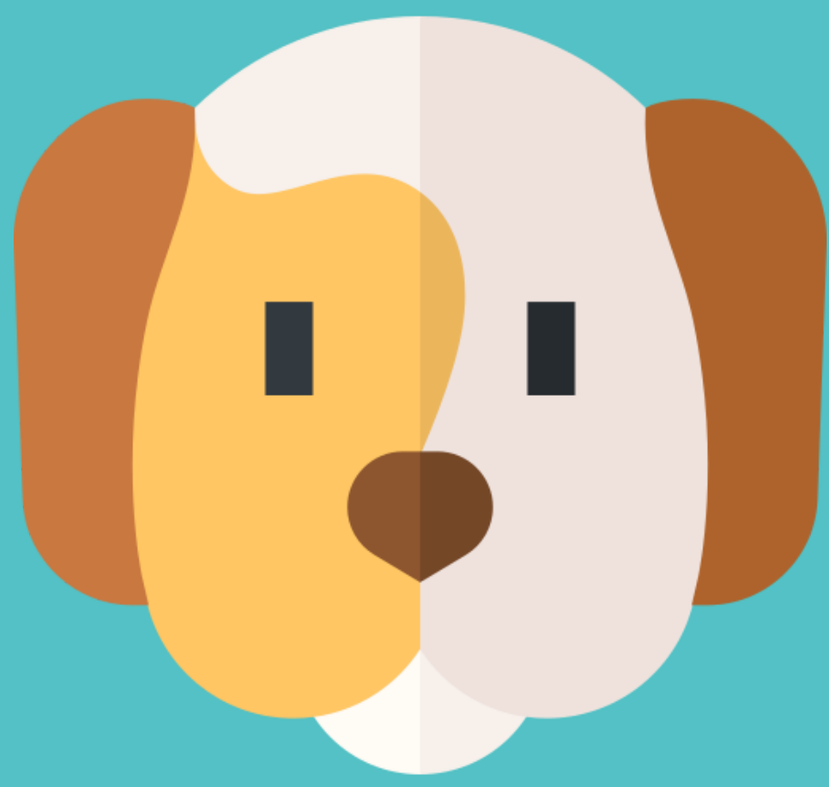
O aprendizado supervisionado é quando temos em nossa base de dados a nossa coluna alvo, ou seja a informação que estamos querendo prever. abaixo temos uma imagem com alguns dados do conjunto de dados do Titanic, que está disponível no site do Kaggle, uma plataforma de competições em Data Science. O objetivo da competição do Titanic é realizar uma predição para saber quem sobreveria ou não ao Titanic. Se prestarmos atenção, o conjunto de dados já contém a coluna Survived que indica quem sobreviveu ou não, sendo 0 para quem não sobreviveu e 1 para quem sobreviveu.

PassengerId	Survived	Sex	Age	Fare
1	0	male	22	7.25
2	1	female	38	53.1
3	1	female	26	16.7
4	1	female	35	26.5
5	0	male	35	16
6	0	male	54	13

Em aprendizado não supervisionado, não teremos a variável alvo em nosso conjunto de dados, ao passar nossos dados para um algoritmo de aprendizado de máquina não supervisionado, o que vai acontecer é que o algoritmo tentará encontrar padrões e criará grupos com quem possui características semelhantes.

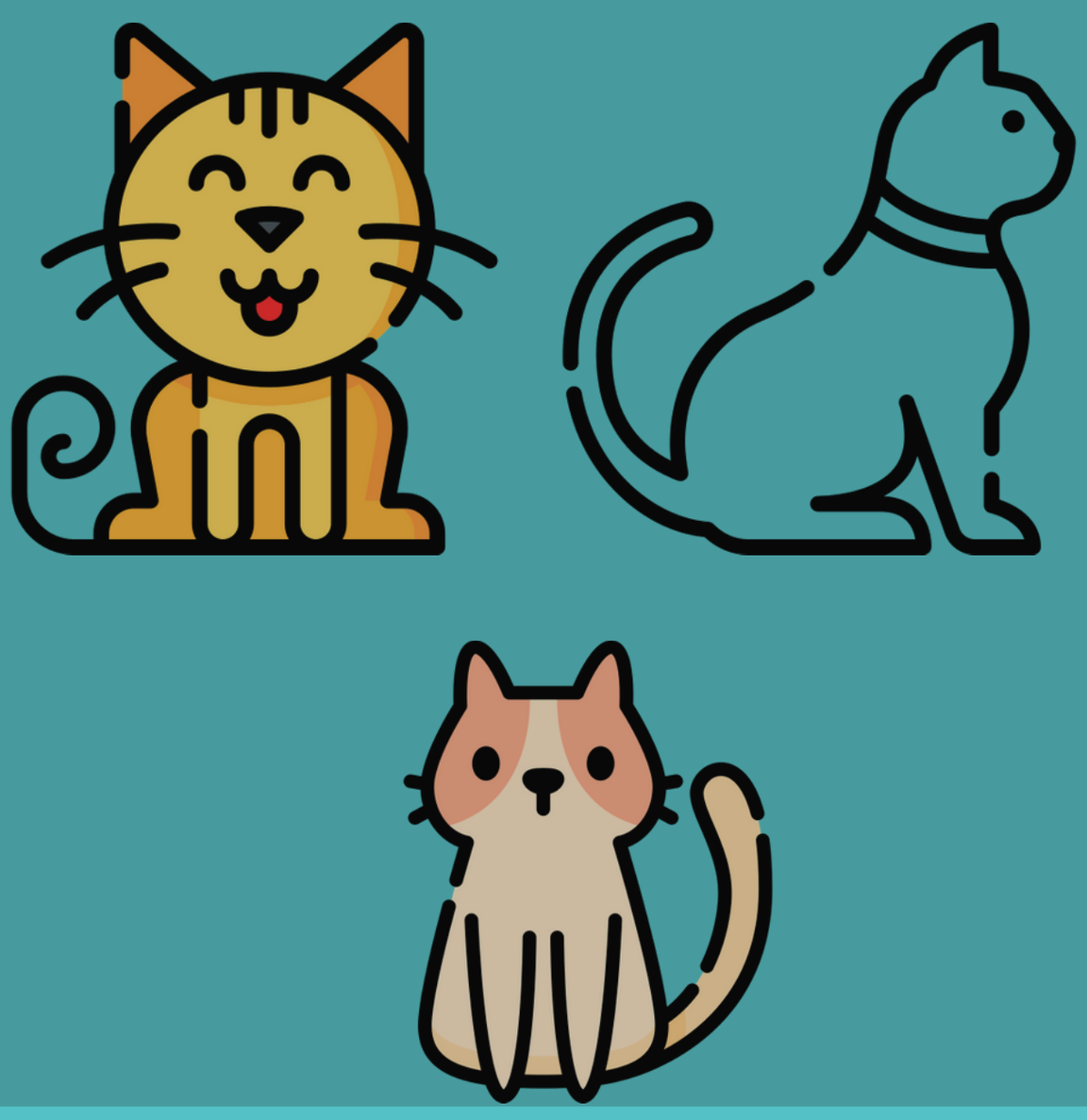
Vamos a um exemplo prático, suponha que temos uma base com imagens de gatos e cachorros, como mostrado abaixo.

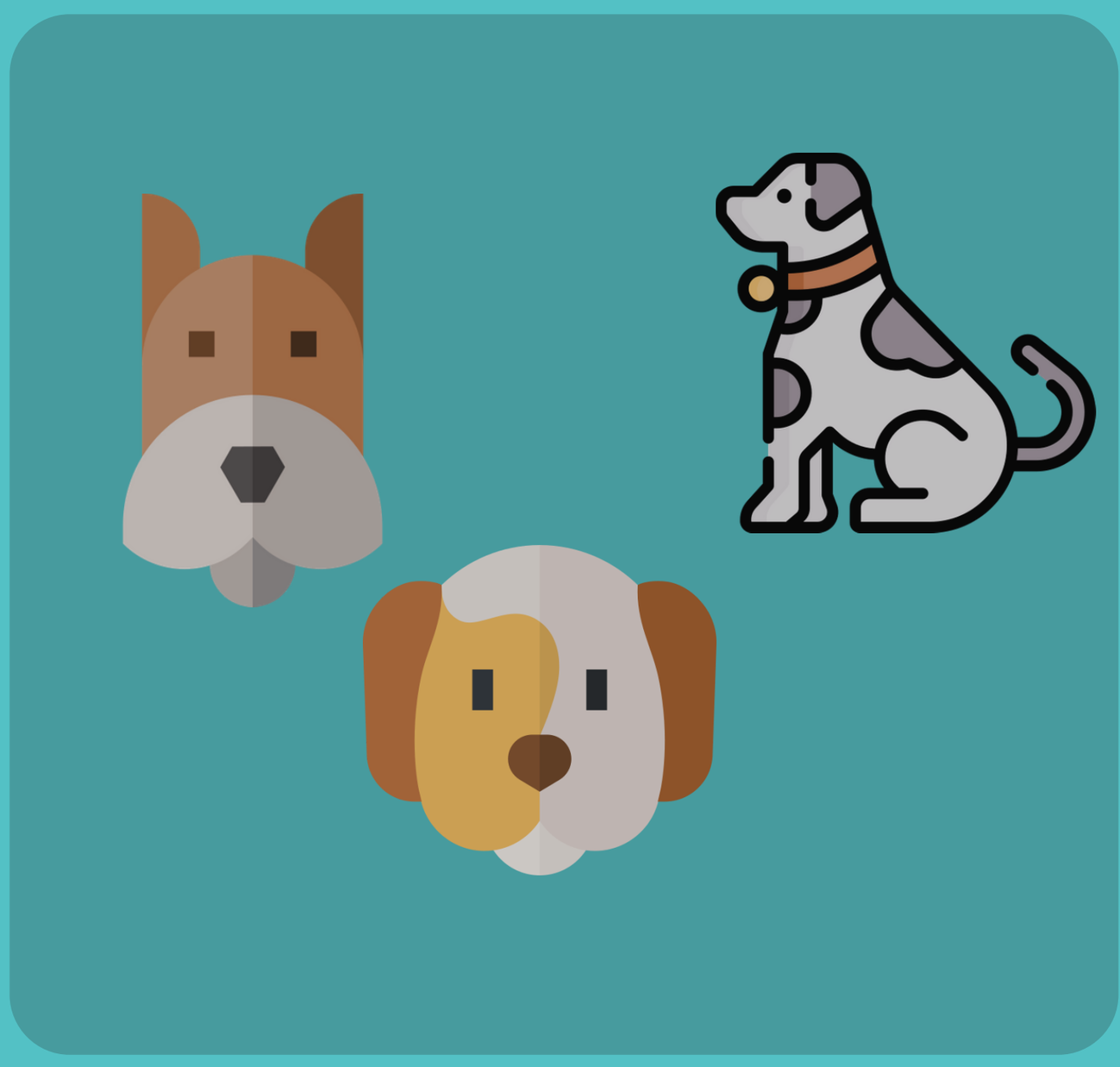




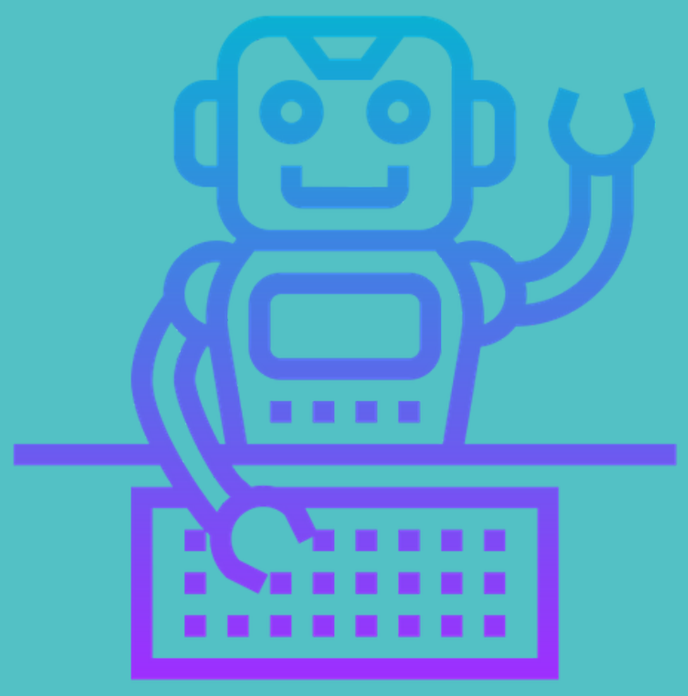
"Ícone criado por Freepik de www.flaticon.com "

Ao passar nosso conjunto de dados com as imagens de gatos e cachorros para um algoritmo não supervisionado a saída será parecida com o mostrado abaixo:





O algoritmo criará 2 grupos, um com as imagens de gatos e outro com as imagens de cachorros, acertará 100%? Muito difícil, mas, quanto mais dados a gente tem, mais o algoritmo aprende e melhora seu desempenho. Agora vamos falar do aprendizado por reforço, o aprendizado por reforço está mais próximo a área da robótica, onde teremos um agente ou mais, e o agente aprende com as interações com o ambiente(Causa e efeito) e também aprende com sua própria experiência. Como exemplo pode-se citar o robozinho que aspira a casa, é um agente que aprende com as interações com o ambiente, neste caso, a casa e que também aprende com sua própria experiência, ao esbarrar em um móvel por exemplo, ele identifica que ali tem um obstáculo e continua a procurar os espaços livres para continuar seu trabalho.



"Ícone criado por Eucalyp de www.flaticon.com "

Considerações Finais

Com a implantação da análise de dados é possível afirmar, por exemplo: qual o produto foi o mais vendido? Em qual período houve queda nas vendas? Quais os produtos que estão com o estoque baixo? São vários os exemplos e não resta dúvidas que uma empresa que aplica o *Business Intelligence* em suas análises, sai na frente com relação a inteligência competitiva, pois, quem conhece suas forças e fraquezas consegue traçar um plano de ação não baseado no 'achismo' e sim na informação que foi gerada, a partir da análise de seus dados.

Contatos

Referências

ALECRIM, Emerson. O que é Big Data? Disponível: https://www.infowester.com/big-data.php. Acesso em: 20 set. 2018.

HOLTZ, Yan; HEALY, Conor. From data to viz. Disponível em: https://www.data-to-viz.com/>. Aceso em 31 ago. 2020.

STEWART, Matthew. O poder da visualização na ciência de dados. Disponível em: <https://towardsdatascience.com/the-power-of-visualization-in-data-science-1995d56e4208 > Aceso em 31 ago. 2020.

TOMÁEL, Maria Inês. SILVA, Terezinha Elisabeth da. A gestão da informação nas organizações. Londrina: Inf .Inf.,v. 12, n. 2, jul./dez. 2007

VISUALIZAÇÕES WTF. Disponível em: < https://viz.wtf/ >. Acesso em 31 ago. 2020.