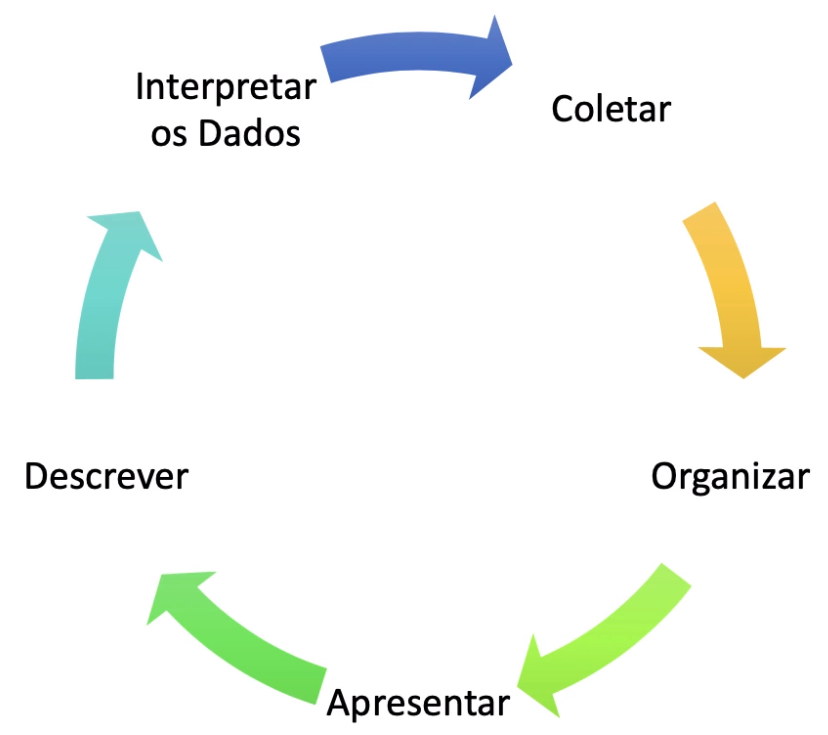
Análise Estatística de Dados

A análise estatística de dados vai nos permitir utilizar os recursos da estatística com uma série de ferramentas para extrair informações relevantes para compreensão da situação atual e para tomada de decisões A possibilidade de analisar os dados à medida que são gerados irá auxiliar e potencializar os resultados empresariais a partir da tomada de decisão.

A estatística em si, é a ciência que nos permite aprender a partir dos dados. Como vivemos na era do Big Data, grandes conjuntos de dados em alto **volume**, alta **variedade** e alta **velocidade**, a estatística se torna uma **ferramenta fundamental** nos dias de hoje na aplicação do fenômeno do Big Data.

Coletar dados: a estatística nos oferece ferramentas como **técnicas de amostragem.** Dificilmente coletaremos **todos os dados** sobre um único fenômeno. As entidades fazem pesquisas com **amostras da população.** As amostras são baseadas em técnicas de amostragens e procedimentos específicos.

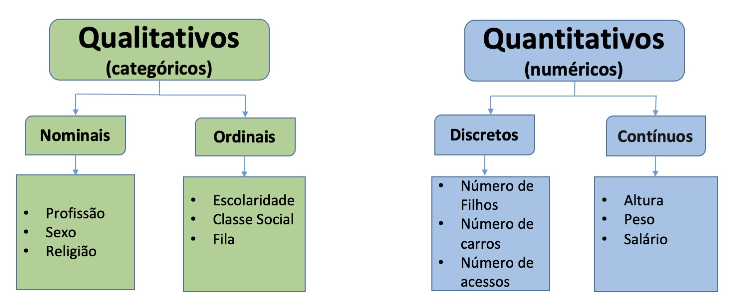
A estatística oferece uma serie de ferramentas para organização de dados, tabulação, calcular frequência, para na sequência poder realizar processo de **análise preditiva.**



Com a estatística também podemos apresentar os dados através de gráficos estatísticos e visualizações que resumem ou simplificam aquilo que nós temos nos dados. Podemos também descrever os dados, ou seja, qual a média do conjunto de dados, qual a mediana, quais os valores extremos do conjunto, a distribuição dos dados. Toda a **descrição dos dados** nos ajuda a compreender como os dados estão organizados para facilitar o trabalho de tomada de decisão.

O trabalho mais importante de todos é a interpretação dos dados através de ferramentas que a estatística oferece. A partir dessa interpretação, podemos fazer **inferências.** Portanto, coletamos amostras, organizar esses dados, descrevê-los e apresentá-los e depois interpretar. Para que por fim possamos fazer **inferências** sobre uma população ou sobre um fenômeno qualquer.

Definindo tipos de dados



Dados qualitativos nominais: representam descrições para os dados e não permitem ranqueamento. Exemplo: CEP (09531-232). Uma informação categórica representada por números.

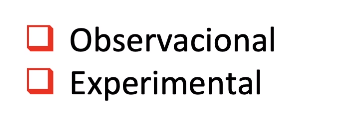
Dados qualitativos ordinais: existe uma ordenação entre as categorias (ranqueamento) e os dados **podem ser medidos.** Exemplo: Graduação e classe social.

Dados quantitativos discretos: valores baseados em observações que **podem ser contados**, normalmente representados por valores inteiros. Exemplo: Número de acesos no sistema.

Dados quantitativos contínuos: valores baseados em observações que **podem ser medidas** e normalmente representados por valores decimais. Exemplo: Salário.

Dependendo do tipo de dado, haverá uma escolha mais aderente ao tipo de técnica estatística aplicada ao conjunto. Se esta definição não for bem feita, o processo poderá ser comprometido.

Qual tipo de estudo estatístico trabalharemos:

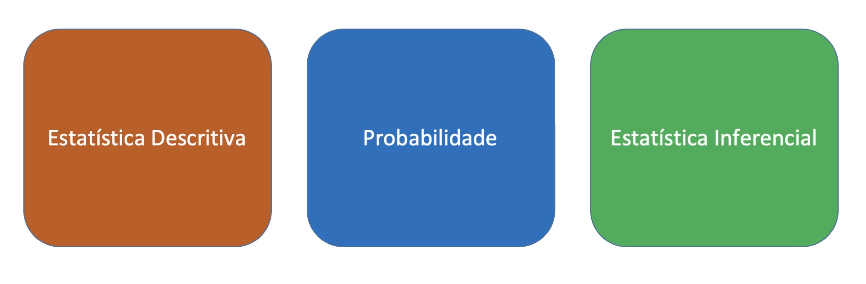


Em um estudo de **observação**, os dados e as características específicas são recolhidos e observados, entretanto, não há iniciativa de modificar os estudos que estão sendo realizados. Neste caso estamos apenas observando o fenômeno, coletando os dados para então realizar o trabalho de análise. Ao analisar dados em tempo real, como coleta de tweets, não se influencia na geração dos tweets, mas apenas **observar** em tempo real, coletar e realizar o processo de análise.



Em um estudo **experimental**, cada indivíduo é aleatoriamente atribuído a um grupo de tratamento, em seguida, os dados e as características específicas são observados e coletados. Os estudos experimentais ajudam a proteger contra potenciais vieses desconhecidos que interferem no resultado da análise estatística, enquanto estudos observacionais não oferecem esse nível de proteção contra **fatores de confusão** na hora de observar os dados. Não existe um estudo melhor do que o outro, cada um será **aderente a um tipo de necessidade** de acordo com o **objetivo da análise**

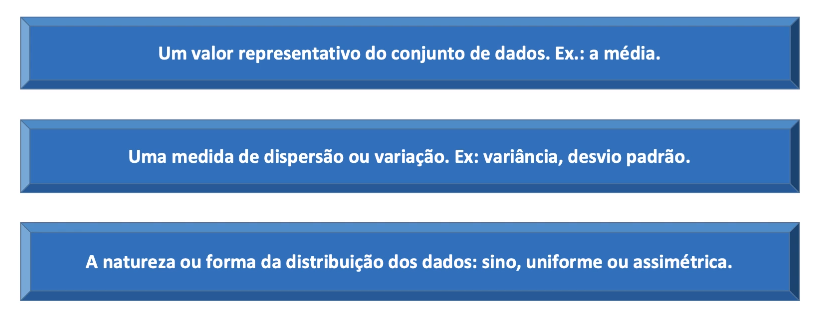
Principais Área da Estatística



Estatística descritiva: É um conjunto de métodos que nos ajuda a descrever os dados, usamos durante o início do processo de análise e depois compreender esses dados, a estatística descritiva oferece uma série de ferramentas para visualização de média, mediana, moda, valores máximo e mínimo, qual forma da distribuição dos dados e uma série de outras informações. É essencial perceber o que **cada método revela** sobre os dados, ou seja, por que descrever os dados? Quais são as informações contidas intrinsecamente no conjunto.

Os dados em si são apenas material bruto, ao descrevê-los é quando conseguimos realmente coletar informações que depois vão se tornar em conhecimento para então auxiliar na tomada de decisão. Em geral a natureza dos métodos descritivos é de **natureza gráfica ou numérica.** A visualização de dados tem papel **fundamental** em todo o processo de ciência de dados.

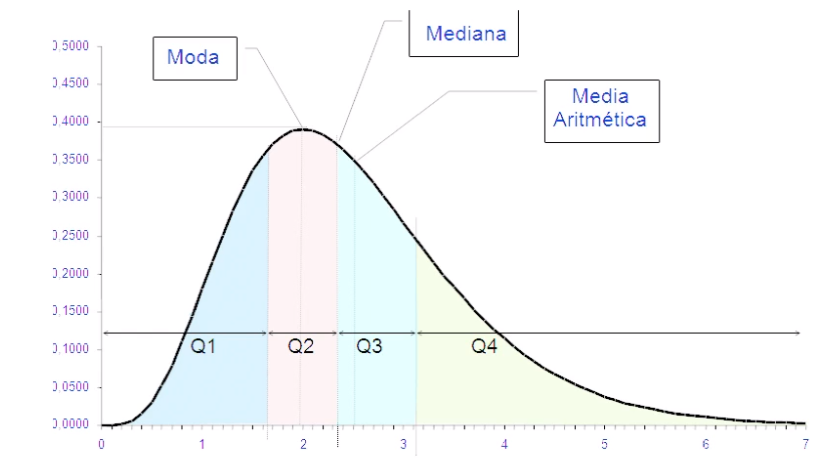
Portanto, estatística descritiva tem por objetivo **sumarizar e mostrar dados**, de forma que se possa rapidamente obter uma visão geral da informação que está sendo analisada. Entendemos melhor um conjunto de dados através de suas três principais características.



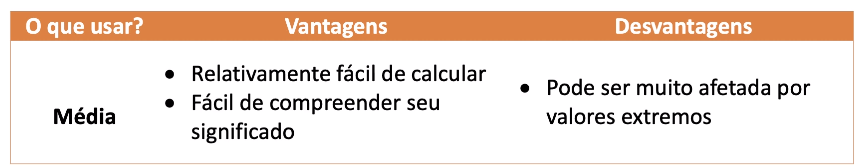
Coletamos os dados e então damos início a análise exploratória, aplicamos estatística descritiva para **obter um valor representativo dos dados**, uma **medida de dispersão e variação** e a natureza ou **forma de distribuição.** Com base na organização dos dados tomaremos decisões seguintes sobre quais ferramentas usamos para tratar os dados, fazer a limpeza, realizar processo de transformação e para então preparar, normalizar e padronizar os dados para modelagem preditiva. As decisões que virão na sequência dependem de informações das **características** do conjunto de dados.

Medidas de Tendência Central nos ajudam a compreender a **centralidade dos dados**

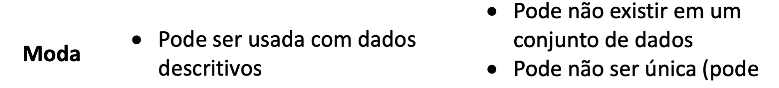
As principais medidas de **tendência central** utilizadas em Estatística Descritiva:



Média Aritmética: é uma medida de tendência central dos dados, ou seja, um número em torno do qual um dataset inteiro está distribuído. É um número único que pode estimar o valor do conjunto de dados completo. São as formas mais simples de identificar tendência em um conjunto de dados. Entretanto, médias podem trazer armadilhas que levam a conclusões distorcidas.



Moda: é o termo que aparece mais vezes no conjunto de dados, ou seja, o termo que tem a frequência mais alta. Ou pode haver um conjunto de dados em que não há nenhuma Moda, pois todos os valores aparecem o mesmo número de vezes. Se dois, três números aparecem o mesmo número de vezes, o conjunto será bimodal, trimodal ou multimodal



Mediana: é o valor que divide os dados em 2 partes iguais, ou seja, o número de termos no lado direito é igual ao número de termos no lado esquerdo quando os dados são organizados em ordem crescente ou decrescente. A mediana será um elemento do **meio da distribuição**, se o número de termos for ímpar. A mediana será a média de 2 elementos do meio da distribuição, se o número de termos for par.



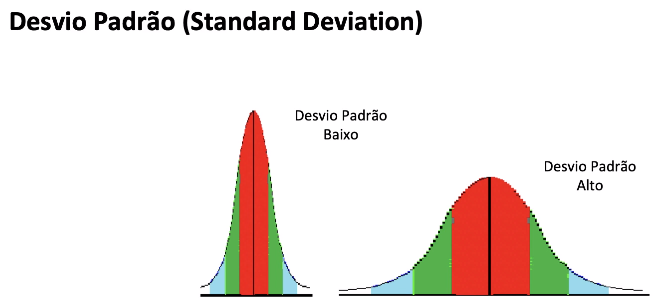
Medidas de Dispersão referem-se à variabilidade dentro do conjunto de dados.

Além de compreender a **centralidade**, precisamos entender a **variabilidade.**

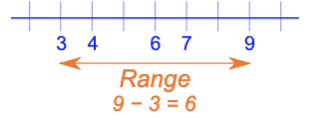
Variância: é o quadrado do desvio padrão. Em algumas situações usa-se um ou outro. A diferença é que com a variância não temos uma unidade.

Desvio Padrão: é a medida de **distância média** entre cada elemento e a **média.** Isto é, como os dados são distribuídos a partir da média. Um desvio padrão baixo indica que os pontos de dados tendem a estar próximos da média do conjunto de dados, enquanto um desvio padrão alto indica que os pontos de dados estão espalhados em uma faixa ampla de valores.

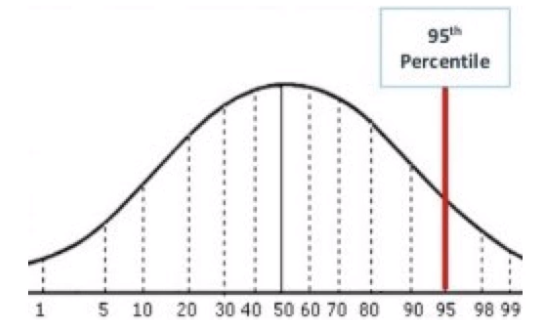
Uma vez que a média é calculada e compreendemos a centralidade dos dados, temos a ideia da distância de cada elemento para esta média.



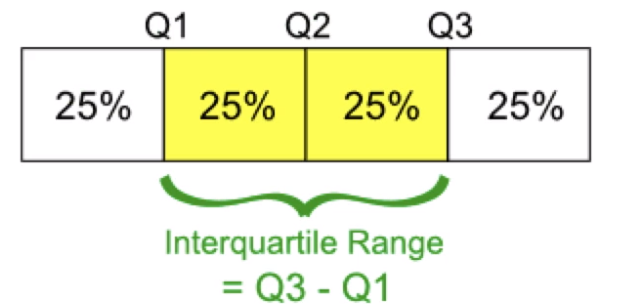
Intervalo: é uma das técnicas mais simples de estatística descritiva. É a diferença entre o menor e o maior valor do conjunto de dados



Percentil: é uma maneira de representar a posição de um valor no conjunto de dados. Para calcular o percentil, os valores do conjunto de dados devem estar sempre em ordem crescente.

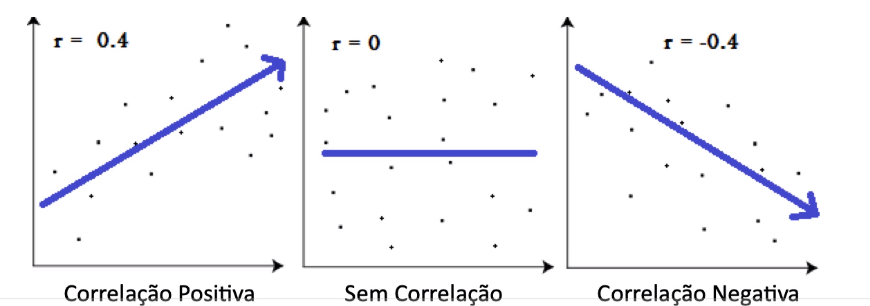


Quartil: são valores que dividem os dados em *quarters,* desde que os dados sejam classificados em ordem crescente. A partir deste cálculo conseguimos identificar em qual quartil os dados estão posicionados



Coeficiente de Correlação: usado com muita frequência durante a fase de **análise exploratória.** Compreender como uma variável se **relaciona com outra variáveis** dentro do conjunto. A correlação permite determinar quão fortemente os pares de variáveis estão relacionados.

O principal resultado de uma correlação é chamado de **coeficiente de correlação** (“r”). Variando de -1.0 a 1.0. Quanto mais próximo “r” for +1 ou -1, mais próximas as duas variáveis estarão relacionadas.

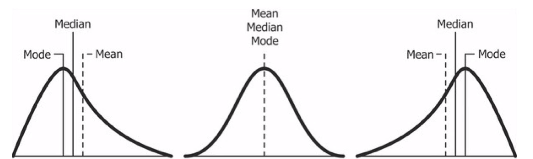


Medidas de forma – Skewness e Kurtosis

As medidas de assimetria (skewness) e curtose (kurtosis) caracterizam a forma de distribuição de elementos em torno da média. Basicamente são dois **coeficientes calculados** para identificar a forma da distribuição dos dados com relação à média. A partir daí fazer inferência, leituras e compreender como os dados estão distribuídos. De acordo com a distribuição dos dados, termos que aplicar alguma técnica para **ajustar os dados** antes da modelagem preditiva. Podemos então calcular o coeficiente e com base no resultado tomamos as decisões, quais técnicas aplicar aos dados para fazer as devidas transformações e ajustes de acordo com o algoritmo de machine learning.

Medida de Assimetria (Skewness): é uma medida de **assimetria da distribuição** de probabilidade de uma variável aleatória de valor real sobre sua média. O valor da assimetria pode ser positivo, negativo ou indefinido.

Em uma distribuição normal perfeita, as caudas de cada curva são imagens espelhadas. Uma distribuição simétrica.



Quando inclinada para a direita, a cauda no lado direito da curva é maior do que a cauda lado esquerda, e a média é a média é maior que a moda. Chamada de assimetria positiva. Talvez seja necessário **aplicar uma técnica estatística** para trazer os dados para uma distribuição simétrica. Alguns algoritmos de Machine Learning exigem essa normalização.

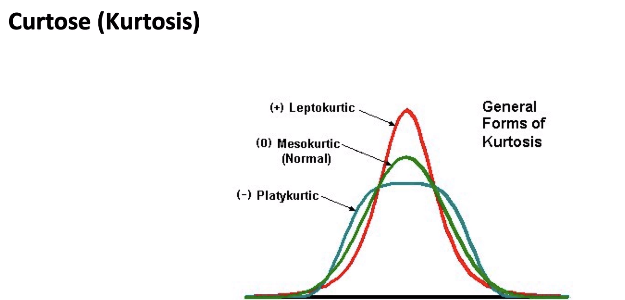
Quando uma distribuição é inclinada para a esquerda, a cauda do lado esquerdo da curva é maior que a cauda do lado direito e a média é menor que a moda. Chamada de assimetria negativa.

Interpretação da assimetria Skewness: a direção da assimetria é dada pelo sinal. Um zero **significa nenhuma assimetria.** Valor negativo significa que a distribuição é negativamente assimétrica. Um valor positivo significa que a distribuição é positivamente assimétrica.

O coeficiente **compara** distribuição da amostra com uma distribuição normal. Quanto maior o valor, **mais a distribuição difere** de uma distribuição normal.



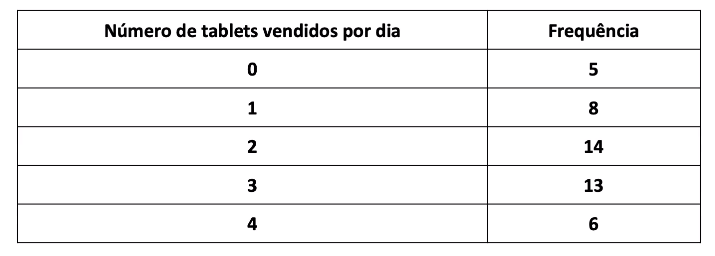
Medida de Curtose: Um dos coeficientes mais utilizados para medir o grau de achatamento ou curtose de uma distribuição é o **coeficiente percentílico de curtose**, ou simplesmente coeficiente de curtose (k), calculado a partir do intervalo interquartil dos percentis de ordem 10 e 90. Curtose é basicamente o contrário do coeficiente de assimetria Skewness. Curtose indica se a curva está mais **para cima ou para baixo.** Mais ou menos achatada.



Se k = 0,263 > distribuição mesocúrtica  
Se k = 0,263 > distribuição platicúrtica  
Se k = 0,263 > distribuição leptocúrtica

Tabela de Frequência

Um dos meios mais simples de descrever dados é através de tabelas de frequência, que refletem as observações feitas nos dados. Observamos um determinado fenômeno, coletamos os dados, e depois tabulamos, criando assim uma tabela de frequência. Relação de cada fenômeno com a frequência correspondente.



Cada linha em uma tabela de frequência corresponde a uma classe. Existem várias ferramentas em R e Python que constroem tabelas de frequência. Esta tabela é tão **importante** **e comum** em ciência de dados e inteligência artificial.

Uma das principais aplicações de inteligência artificial na atualidade é o processamento de linguagem natural, para sermos capazes de uma aplicação de computador, reconhecimento de voz, tomar uma ação ou gerar resumos automáticos. Uma das primeiras etapas no **processamento de linguagem natural** é criar a tabela de frequência com a ocorrência de cada palavra em um texto.

Distribuição de Frequência: mostra o **número de observações** de dados que estão em um intervalo específico. É basicamente uma forma de colocar mais informação em uma tabela de frequência

Como construir uma Distribuição de Frequência?



Com a distribuição de frequência construída, podemos coletar uma série de insights sobre os dados, ou seja, primeiro **construímos a distribuição de frequência** para então poder compreender melhor os dados.

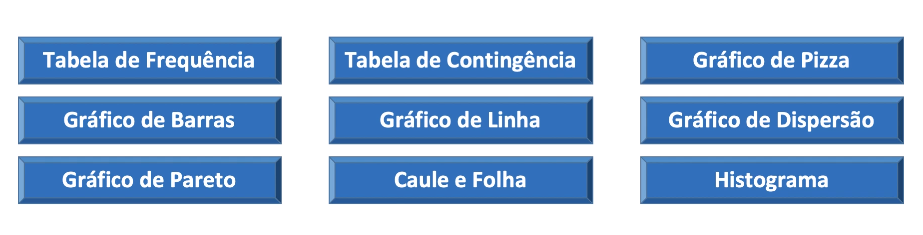
Principais ferramentas da Estatística Descritiva

Tabela de frequência basicamente mostra a ocorrência dos elementos dentro do seu conjunto de dados.

Tabela de contingência é usada principalmente quando nós temos duas variáveis e desejamos ver a relação entre as duas.

Gráficos como ferramenta de análise para compreender como os dados estão organizados, distribuídos e como se relacionam.

Histograma mostra de maneira visual a frequência dos dados

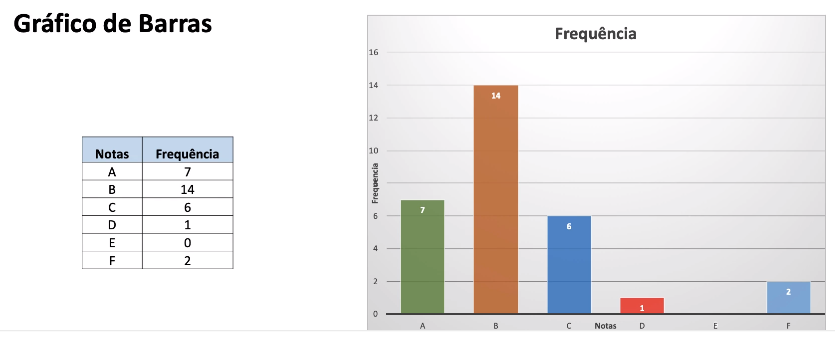


Temos duas classificações principais para as ferramentas oferecidas pela Estatística Descritiva



Gráfico de barras é um dos gráficos mais utilizados em análise de dados. A tabela de frequência é representada através de barras em um gráfico de barras. Cada barra representa exatamente a proporção da frequência na tabela de frequência.

O principal ponto com relação a este gráfico é a escala. A escala é uma das principais formas de se manipular um gráfico destes, para que um gráfico represente a informação que interessar.



O gráfico de Pareto pode ser construído através das barras representando cada uma das classes na tabela de frequência. A altura das barras está diretamente associada à frequência e ao lado direito o percentual. A linha que passa por todo o gráfico é construída de modo que do lado esquerdo da linha se tenha a principal causa do problema e do lado direito as causas menos relevantes para o problema. Rapidamente é possível identificar as causas menos relevantes para os problemas. Muito usado em **controle de qualidade**.

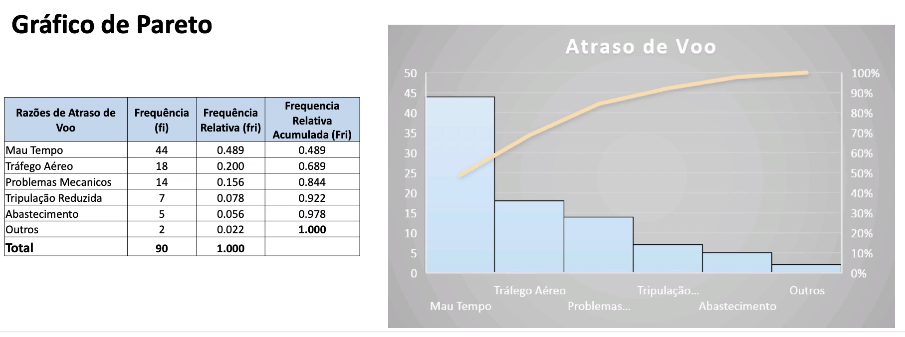
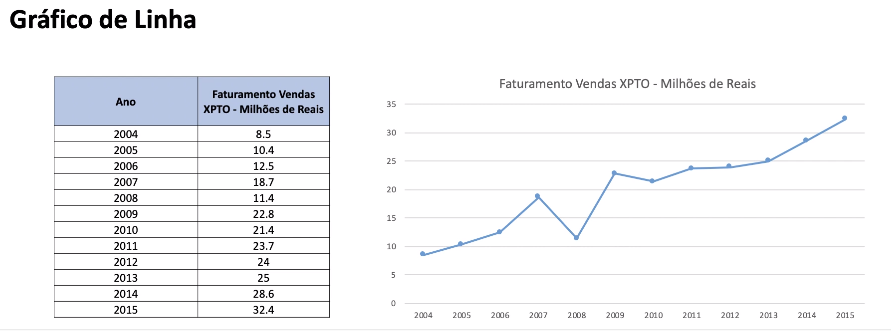
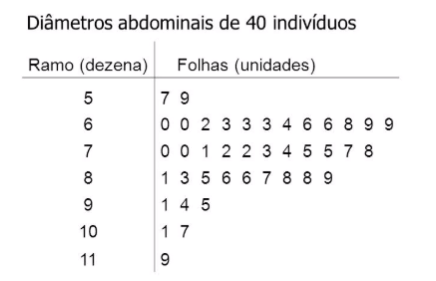


Gráfico de linhas é amplamente usado na análise de dados, basicamente usado para retratar a evolução de uma variável ao longo do tempo.

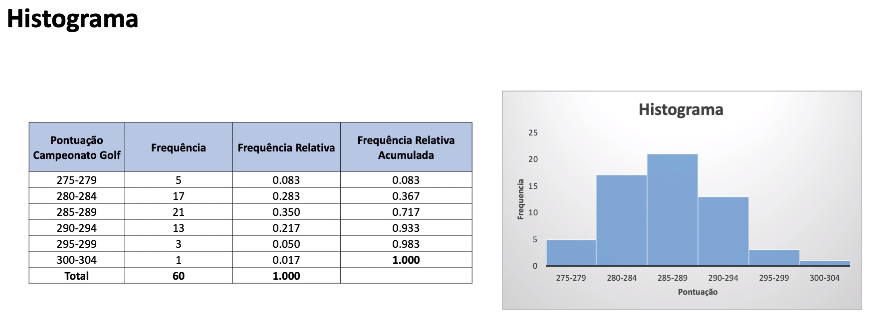


O gráfico Caule e Folha, divide os dados em duas partes: O **caule** (ramo) representa os valores maiores e ficam à esquerda do traço vertical. As **folhas** são os menores valores, ficam à direita do traço vertical. Listando todas folhas à direita de cada caule, podemos graficamente descrever como os dados estão distribuídos.

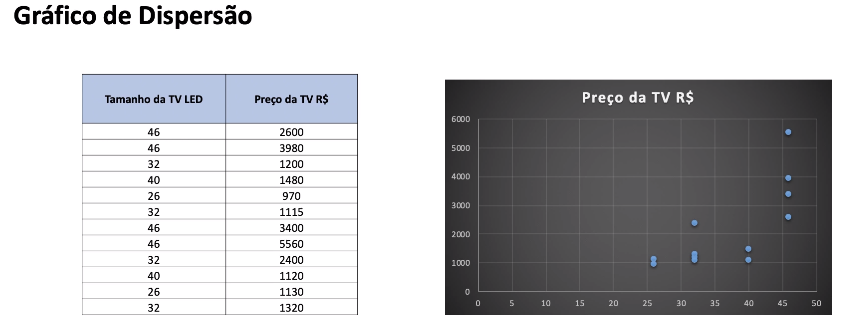


Ferramenta que será usada constantemente, o **histograma**, tem aparência muito próxima ao gráfico de barras. A informação contida nele ou interpretação é em geral diferente. A principal intenção no uso do Histograma é mostrar a **distribuição de frequência** e com isso analisar se os dados seguem ou não uma distribuição normal. A forma do histograma diz muito sobre forma em que os dados estão organizados ou distribuídos, em geral, antes de iniciar o processo de **modelagem preditiva**, o histograma é usado antes do pré-processamento.

Dependendo do algoritmo que formos usar mais a frente, é necessário mudar a distribuição dos dados **aplicando a normalização**.



Ferramenta usada para análise **bivariada.** O objetivo é olhar para duas variáveis e como se relacionam.



Mostra a relação numérica entre duas variáveis. O objetivo é resumir as variáveis.



Probabilidade: base de diversos algoritmos de aprendizado de máquina. Temos uma categoria inteira de **algoritmos de machine learning** que são algoritmos probabilísticos. Portanto compreender a teoria da probabilidade ajuda não apenas a analisar dados, mas também compreender vários algoritmos de aprendizado de máquina.

Estatística inferencial: focada em previsões, normalmente se trabalha com amostras de dados para fazer inferências sobre uma população, fenômeno ou assunto qualquer. A estatística inferencial oferece uma **série de teorias e ferramentas** que estão por trás da grande maioria dos algoritmos de machine learning. Além de ser usada para análise, também é necessário compreender para aplicação de machine learning.