

ANÁLISE DESCRITIVA



EJ ANALYTICS

AGENDA

01

Introdução

02

Estatística

03

Análise Descritiva

04

Testes de Hipóteses

05

Limpeza de dados





1. Introdução

01 Introdução

Esse treinamento busca **ensinar principalmente a estatística** necessária para fazer a **análise de dados**, abordando principalmente um **nível mais avançado** para tal. Aqui, os **termos mais teóricos** são ensinados para que seja possível entender bem a **teoria** por trás de um **modelo de classificação** e aplicá-lo na **prática**.

Além disso, também são evidenciadas **técnicas estatísticas** para se **analisar dados** junto com **testes de hipóteses** para testar se esses valores estão realmente bem adequados dentro de um possível **modelo realizado**.

Por fim, é mostrado um **treinamento de limpeza de dados no Excel** utilizando uma **Pesquisa Quantitativa** e o projeto da **TOK&STOK** como base, para que se possa extrair diferentes **insights** após a base estar bem organizada para isso.





2. Estatística

02 Estatística

A **análise descritiva**, como o próprio nome já diz, refere-se a **analisar os dados** em **linhas gerais**, de modo a entender **características qualitativas** de **amostras quantitativas**. Primeiro, deve-se diferenciar o que é uma **amostra** do que é uma **população**:

População: são todos os **dados disponíveis** de uma determinada **variável**. Por exemplo, todas as alturas das pessoas no mundo representam uma população da variável chamada de alturas. Os **dados financeiros** normalmente se encaixam aqui, pois você possui **todos os dados existentes** daquela variável (ex: preços de ações).

Amostra: é uma **parte particionada de uma população**. Por exemplo, as alturas dos **membros da EJ** são uma **amostra** da variável alturas. Normalmente, queremos utilizar amostras para **descobrir alguma característica específica** de uma população e criamos **hipóteses** para dizer isso.



02 Estatística

Exemplo: em uma **pesquisa quantitativa**, assumimos que aquela base de dados foi selecionada por uma **amostragem aleatória simples (AAS)**, ou seja, os dados foram obtidos de **modo aleatório e com reposição**, para que as características da amostra se pareçam com as **características** da população.

Porém, resta também entender como são as **características dessa população**. Para isso, são calculados os **momentos**, que são **quantitativas** que exprimem algum significado da **amostra**. São eles: **média, variância, correlação, simetria e curtose**. Os dois últimos não são tão utilizados e outra ferramenta além dessas também é muito **utilizada**, que são os **quartis**.



02 Estatística

Esses momentos servem para tentar **adivinhar algo da população**. Para isso, utilizamos **estimadores**, que são **funções** das **amostras** observadas e, caso haja uma **AAS**, existe uma **certa probabilidade** de essas estatísticas se **aproximarem** do valor **populacional**.

Então, por exemplo, lidando novamente com **alturas**, pegamos **vários grupos aleatórios de alturas**, alturas da EJ, da Poli JR, da Insper JR, da ESPM Jr. Em cada um desses grupos, podemos **calcular a média dessas amostras**, e esse estimador é chamado de **média populacional**.

Porém, quando há **AAS**, a **média da média amostral** é igual à **média populacional**, ou seja, se retirarmos a **média das médias de altura** de todos esses grupos, ela tenderá a se **aproximar para a média populacional**. Nesse sentido, a média amostral é um bom **estimador** para entender **médias populacionais** em amostras aleatórias.





3. Análise Descritiva

03 Análise Descritiva

Média: é o **valor central de uma distribuição**, servindo muito como um **indivíduo representativo** daquela amostra. Como já foi dito, a **média amostral** acerta em média o seu valor populacional, o que significa que ela é um **estimador não viesado** da média populacional (**acerta em média**). A **variância** da média amostral é a **variância da amostra dividida por n** (quanto mais dados, mais a distribuição tende para a média pois a variância tende a zero).

O problema da média é que ela é **muito sensível a outliers**, então elementos muito distantes do restante da amostra podem dificultar a visualização.

Quartis: mostram como os valores que de fato são da amostra são distribuídos e, por isso, são uma medida mais **resistente a outliers**, pois olham os dados pela **posição** e não pelo **valor médio**. Colocando a amostra em **ordem crescente**, temos:



03 Análise Descritiva

Primeiro Quartil: é o valor que deixa **25% dos dados para trás dele**.

Segundo Quartil (mediana): deixa **50% dos dados para trás dele**.

Terceiro Quartil: deixa **75% dos dados para trás dele**.

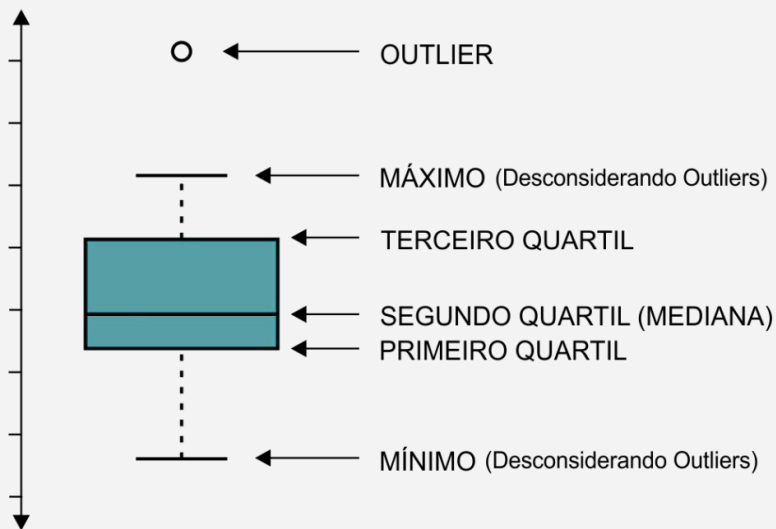
Distância interquartil (dq): é a **distância** entre o **primeiro** e o **terceiro quartil**, representando uma medida resistente do desvio padrão.

Limites superior e inferior: são definidos como **valores acima** do **terceiro quartil** em **1,5 dq** ou abaixo do **primeiro quartil** em **1,5 dq**. Esse valor de **1,5** é **definido** baseando-se na distribuição normal, pois, nela, esse valor de 1,5 faz 1% da amostra ser considerada como **outlier**.



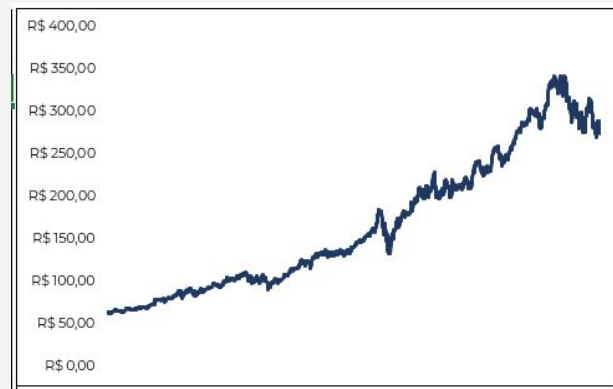
03 Análise Descritiva

Outliers: são valores que **passam** do **limite superior** ou do **limite inferior** da amostra, então são calculados a partir dos limites superiores e inferiores e não o contrário. Eles normalmente **devem ser retirados da amostra pois atrapalham a interpretação** dos resultados.

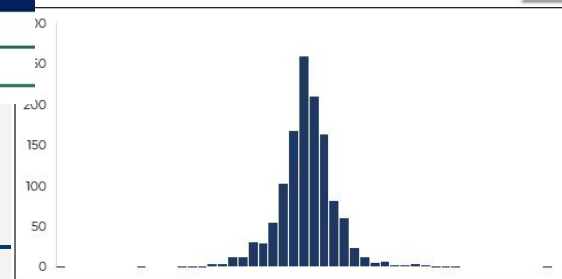
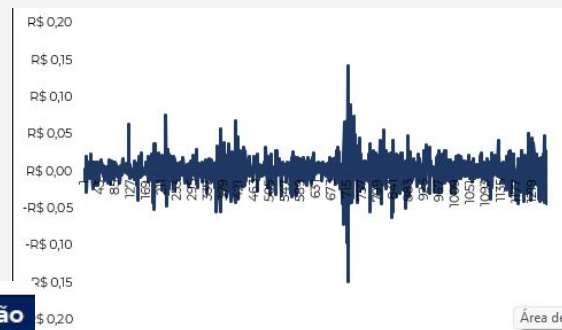


03 Análise Descritiva

No Excel, há um **pequeno roteiro para remover outliers**. Primeiro, **cria-se o gráfico** da série e, se ela **tiver uma tendência** (crescer ou decrescer com **coeficiente angular diferente de zero**), deve-se **retirar os retornos da variável** ($x_t / x_{(t-1)} - 1$). Depois disso, pode-se criar um **gráfico de linha**, um **histograma** e um **boxplot** para os retornos para observar se ela parece com **desvios** (gráfico na direita).



Índice	Valor	Variação
1	R\$ 64,35	
2	R\$ 64,45	$=F4/F3-1$
3	R\$ 64,55	



03 Análise Descritiva

Depois disso, o *boxplot* vai **mostrar se a série possui ou não outliers**. Caso esteja, deve-se **descobrir quais dados são outliers** utilizando a fórmula dita anteriormente e, caso tenha *outliers*, deve-se **removê-los**. Isso pode ser feito utilizando o **filtro** para descobrir os *outliers* e **depois remover todas as linhas que os tenha** (vale fazer uma cópia para não deletar a variável original). Depois disso, basta **analisar as séries** com o que foi dito anteriormente.

Valor	Variação	Outlier?	Primeiro Quartil	-0,66%
R\$ 64,35			Terceiro Quartil	1,03%
R\$ 64,45	0,16%	$=SE(G4<SR$2-1,5*$	IIQ	Série1 Ponto 19 Valor: R\$ 0,02
R\$ 64,70	0,39%	SR4;1;SE(G4>SR$3+1,$		
R\$ 63,91	-1,22%	$5*SR$4;1;0))$		
R\$ 63,83	-0,13%	0		

03 Análise Descritiva

Variância: é uma **medida** de **dispersão** que mostra como os dados estão **distribuídos**, sendo uma medida de **volatilidade** e risco. Ela é calculada a partir da **média dos desvios à média ao quadrado de uma amostra**, porém, como essa medida está ao quadrado, normalmente interpreta-se dados com base no **desvio padrão**, que é a **raiz quadrada da variância**.

Em **variância populacional**, calcula-se com o denominador **n** (quantidade de elementos), porém, em **termos amostrais**, o **estimador não viesado da variância** é aquele dividido por **n-1**, então este que normalmente é utilizado na estatística.

Correlação/Covariância: é uma medida de **comparação** entre **duas bases de dados** e serve para mostrar o quanto elas **andam juntas**. Algo importante é que a **covariância** e a **correlação** mostram o **grau de semelhança** entre as amostras e por isso é muito importante, sendo que a **correlação** é mais **interpretável** (vai de -100% a +100%).



03 Análise Descritiva

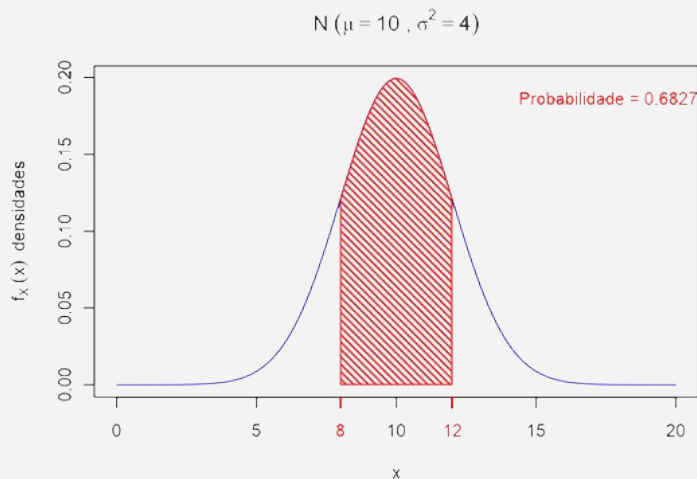
Importante ressaltar também que **correlação não implica causalidade**, pois muitas vezes existem **outras variáveis causando aquele efeito** observado na **correlação** e não correlação não implica **independência**, pois a correlação não capta **efeitos não lineares** nas variáveis.

Em **regressões lineares (tipo de modelos de classificação)**, busca-se suprimir esse **efeito não observado** nas variáveis e, por elas, tenta-se encontrar o **efeito causal** de uma variável na outra, algo mais forte do que uma **simples análise de correlações**. Essas regressões retiram os **efeitos** das outras variáveis **adicionando elas na equação**, porém isso vai ser abordado no treinamento de modelo de classificação.



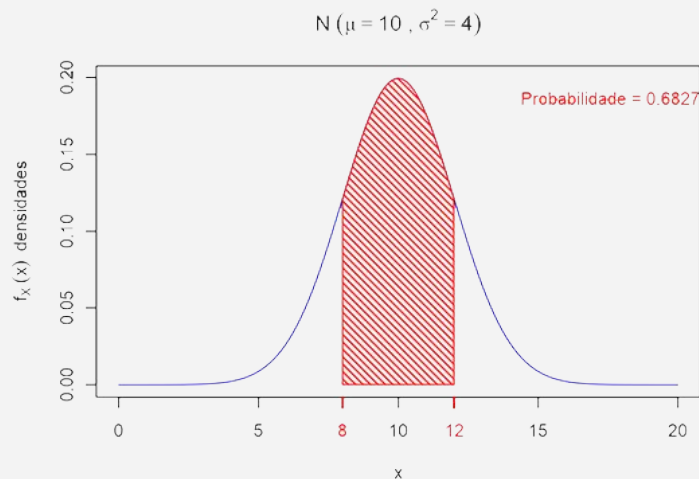
03 Análise Descritiva

Distribuição: é a forma com que os **dados estão distribuídos**, muito parecido com um **histograma**, porém muito mais elaborado. Nesse sentido, o **eixo x mostra quais são os dados** e o **eixo y mostra a probabilidade desse dado aparecer** (ou quantidade que ele aparece). Por exemplo, se a **quantidade de produtos vendidos por semana** tiver a seguinte **distribuição**, então a **probabilidade** de vender **8 produtos é cerca de 12,5%** (olha o eixo y).



03 Análise Descritiva

Distribuição acumulada: é como uma **soma das probabilidades** ou uma **probabilidade acumulada**. No gráfico, é a **área vermelha** e ela mostra a probabilidade de **vender de 8 a 12 produtos** em uma semana (68%). Normalmente se utiliza um **histograma** para fazer esse gráfico, mas também pode ser utilizado o **Python** para fazer algo mais complexo (**código no Drive**).



03 Análise Descritiva

Existem **algumas distribuições** bem **conhecidas** que são **importantes** de conhecer pois aparecem bastante. São elas: **uniforme, normal, qui-quadrado, t de Student e F de Fisher.**

Uniforme: todos os **dados ocorrem** com **igual probabilidade**. Seu gráfico de **distribuição** é uma reta **horizontal**. Se há n ocorrências, então cada uma tem **probabilidade $1/n$** .

Normal: é a **probabilidade** mais **importante** de todas. Toda **normal** é definida por sua **média e seu desvio padrão** e sua cara é muito **característica**: totalmente **simétrica** e com dados que **ocorrem mais perto da sua média**. Ela é a mais utilizada pois consegue abordar muitos dos dados devido à sua **forma simétrica e concentrada no centro**. Alturas se encaixam muito em normais. A **normal padrão** é aquela com **média 0** e **desvio padrão 1**.



03 Análise Descritiva

Qui-Quadrado: é assim chamada pois é a soma de **r normais padrões ao quadrado** e é **um pouco diferente** da **normal**, pois é **assimétrica**, com um pico mais **concentrado**. O valor de **r** é definido como os **graus de liberdade da distribuição** e graus de liberdade são os **parâmetros** dela, alterando sua distribuição conforme eles mudam.

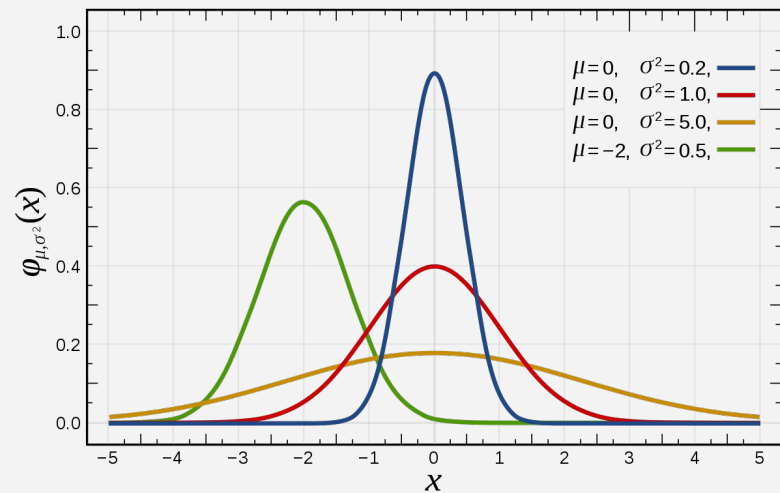
T de Student: tem a mesma cara porém só é definida com base em seus **graus de liberdade**, pois são definidas com base na **qui-quadrado**. A **distribuição t** explica melhor as **situações reais** pois suas caudas são um pouco diferentes da normal, então às vezes explica melhor e também é usada em **testes de hipótese**. Ts com mais de **30 graus de liberdade** tendem a normais.

F de Fisher: também tem a mesma cara que a **qui-quadrado**, porém é definida com base em **dois graus de liberdade**, pois é definida como a **divisão entre duas Ts de Student**. Existe um **teste de F**.

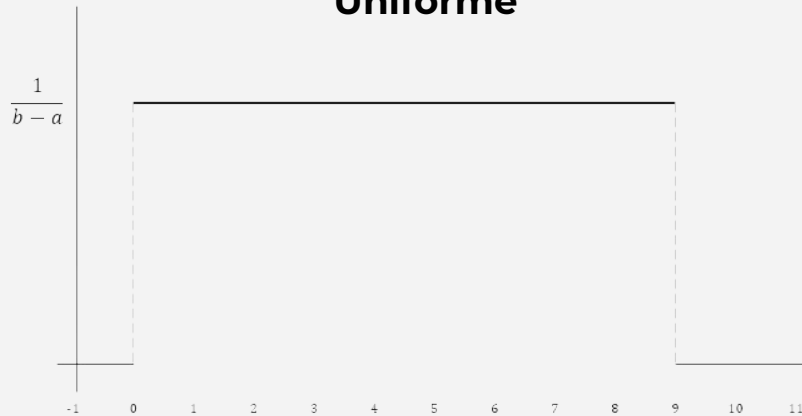


03 Análise Descritiva

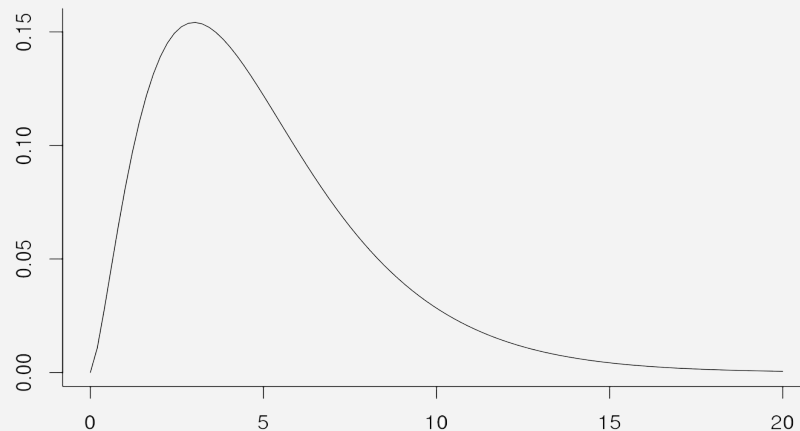
Normal



Uniforme

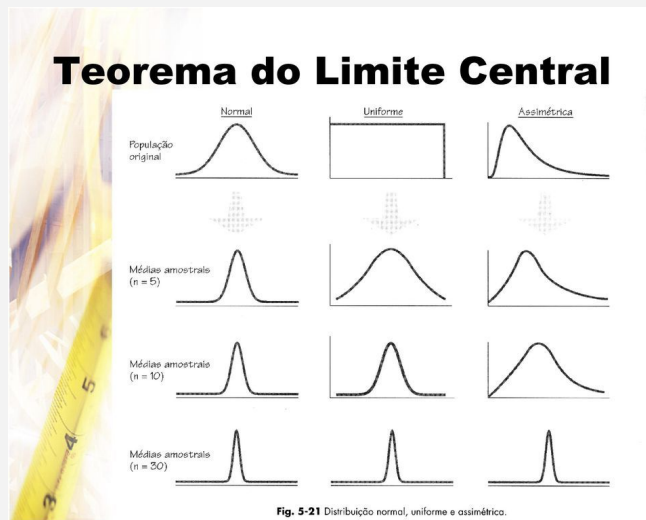


Qui-Quadrado



03 Análise Descritiva

Teorema do Limite Central: diz que a **média amostral**, quando padronizada da **maneira correta**, tende para uma **distribuição normal**. Por isso, muitas das **distribuições amostrais** tendem à normal com **média igual à média da amostra** e **desvio padrão igual ao desvio padrão da amostra dividido por raiz do número de elementos**, pois muitas distribuições surgem da média amostral.





4. Testes de Hipóteses

04 Testes de Hipóteses

Os **Testes de Hipóteses** servem para **testar alguma hipótese**. Na prática, se você quer saber se a **média populacional é igual a algum valor**, você utiliza sua amostra para **testar** se isso é realmente verdade. Para isso, estimadores são necessários para estimar se isso é de fato verdadeiro.

Há diferentes tipos de **testes de hipóteses** e cada um deles exige um **estimador** diferente. É importante entender a **lógica geral dos testes** para saber aplicar em cada um.



04 Testes de Hipóteses

Para fazer um **teste de hipótese**, precisamos dos seguintes **termos**:

Hipótese nula: é o que **queremos testar**, por exemplo $H_0: \mu = 0$ (média igual a zero).

Hipótese alternativa: a opção caso o **teste seja falso**, pode ser **unilateral** ($H_a: \mu > 0$) ou **bilateral** ($H_a: \mu \neq 0$) e unilaterais são mais fáceis de serem rejeitados, pois diz qual direção que a variável deve seguir.



04 Testes de Hipóteses

Nível de Significância ou erro tipo 1 (α): é a **probabilidade** de **rejeitar a hipótese nula** dado que ela é **verdadeira**, ou seja, é um valor definido como uma **possível discrepância** e normalmente é definido como **5%**.

Poder do teste: é a **probabilidade de rejeitar a hipótese nula** dado que ela é **falsa** e ela mostra quanto o **teste aceita quando vai rejeitar uma hipótese**. Quando a **variância** do **estimador** é muito alta, há uma **perda de poder do teste**.



04 Testes de Hipóteses

Nesse caso, se os **dados populacionais** se comportarem como uma **distribuição normal** (como mostrado abaixo), padronizamos ela retirando a média e dividindo pelo desvio padrão, **sua média é a média populacional** e a **região crítica é baseada no nível de significância** definido. Então, caso o erro tipo 1 seja de **5%**, a probabilidade 5% antes definida é a **área pintada de azul**, sobrando **95%** para a região de **aceitação**.



04 Testes de Hipóteses

Como não temos acesso ao **valor populacional**, testamos se nossa **média amostral** está dentro da **região de aceitação**. Para isso, calculamos os valores de z_1 e z_2 a partir dos valores de 5% do nível de significância e, como há 5% na região azul, então há 2,5% para a direita e 2,5% para a esquerda.

E, ao calcular isso, deve usar a **normal acumulada inversa padrão** ao nível de $1 - 5\%/2 = 97,5\%$ e, para a **normal**, esse valor é de **1,96** (vale decorar esse valor). Se a média amostral estiver dentro da região de aceitação (entre z_1 e z_2), aceitamos a **hipótese nula** e, se não estiver, **rejeitamos**.



04 Testes de Hipóteses

Porém, existe outro modo de **testar a hipótese nula**. Podemos **calcular** um **intervalo de confiança** a partir da **média amostral**, que é aquele **intervalo** que **aceitamos a hipótese nula** se ela estiver dentro dele. Para formar isso, **pegamos a média amostral** e calculamos os **limites superior e inferior** a partir da **margem de erro** (média + margem de erro e média - margem de erro).

Caso a **média populacional** definida na hipótese nula esteja **dentro do intervalo de confiança**, aceitamos a **hipótese nula** e, se não estiver, **rejeitamos**.



04 Testes de Hipóteses

Pelo **Teorema do Limite Central**, assumimos que a **amostra** se comporta como uma **normal**. Para criar o **intervalo de confiança**:

Calculamos z com base no **nível de significância** (se for de 5%, então $z = 1,96$) e **multiplicamos ele pelo erro padrão** da nossa média amostral. Porém, a **variância da média amostra** é a variância **dividido por n** (tamanho da amostra), então o erro padrão é a raiz quadrada disso:

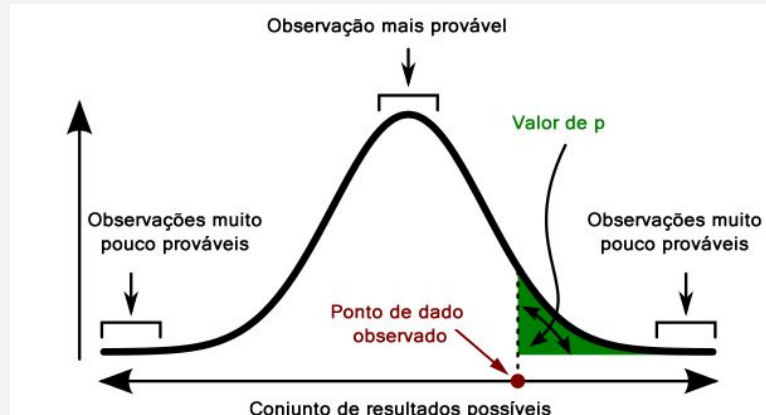
$$\text{LIMITE INFERIOR: } \bar{X} - z \times \frac{\sigma}{\sqrt{n}}$$

$$\text{LIMITE SUPERIOR: } \bar{X} + z \times \frac{\sigma}{\sqrt{n}}$$

04 Testes de Hipóteses

Outro modo de verificar se um **teste de hipótese está ou não correto é pelo p-valor**. O **p-valor** é a **probabilidade** de conseguir um **valor mais extremo** do que o calculado pela **amostra**. Portanto, ele é a **região verde** mostrada na figura.

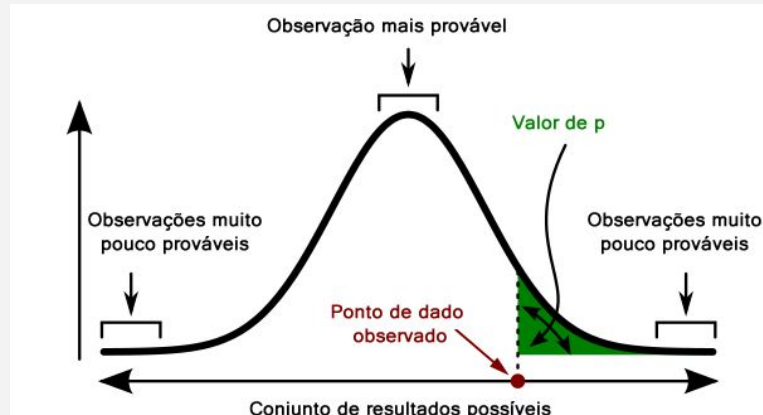
Caso o **p-valor** seja maior que o **erro tipo 1**, significa que nossa **média** está dentro da **região de aceitação**, então aceitamos a **hipótese nula**. Caso o **p-valor** seja menor que o **nível de significância**, **rejeitamos** a hipótese nula.



04 Testes de Hipóteses

Algo importante é que, em **testes bilaterais**, o **p-valor é multiplicado por 2**, pois ele engloba as **caudas da direita e da esquerda** já que a estatística de teste pode ser **positiva** ou **negativa**, o que não ocorre em testes unilaterais.

Dito isso, **p-valor em testes unilaterais é menor** e portanto é **mais fácil de rejeitar** a hipótese nula.



04 Testes de Hipóteses

Teste sobre média com variância conhecida: a **estatística de teste** é a **média amostral**, como definida anteriormente. Esse é o **caso mais simples** em que se aplica tudo dito anteriormente, com **z dos intervalos de confiança** calculados pela normal e **p-valor** calculado com **probabilidade normal**.

Teste sobre média com variância desconhecida: esse é o **caso mais comum** no mundo real e tem apenas algumas diferenças do teste anterior. Nesse caso, **a estatística de teste é uma T de Student**, mas tudo dito anteriormente **funciona bem**, porém, há a diferença de que os **z's calculados anteriormente são agora calculados com uma T inversa acumulada** e o **p-valor** também é calculado por uma **T** (há a opção disso no Excel).

Nessa **T de Student**, há **n-1** graus de liberdade.



04 Testes de Hipóteses

Teste sobre variância: esse teste **não é tão comum**, mas utiliza-se a distribuição **qui-quadrado** para testá-lo. A grande diferença é que o **intervalo de confiança muda**, mas pode-se achar ele **facilmente na internet**.

Teste sobre diferentes médias: esse é o caso **mais completo** e trata de um teste sobre mais de uma **média de distribuições diferentes**. Normalmente calcula-se das **médias serem conjuntamente iguais a zero** (esse é o teste calculado depois de uma **regressão múltipla**) e os **graus de liberdade são $n-1$ de cada uma das distribuições** (com uma diferença para regressões). Não é tão importante entender isso, é mais importante entender o que ele significa.



04 Testes de Hipóteses

Na prática, basta primeiro calcular a **média amostral**, o **desvio padrão** e o **valor de n** (**número de elementos**). Depois disso, deve-se **calcular o valor de z** com base em **qual teste vai ser feito** (como o desvio padrão normalmente é desconhecido, normalmente se calcula ele **com base na estatística T** (como n é grande, o valor também é 1,96).

Assim, basta calcular os **intervalos superior e inferior** por meio da fórmula do **intervalo de confiança** e, para verificar se está certo, pode-se utilizar a **fórmula de intervalo de confiança do Excel**, que calcula a **margem de erro**.

Média amostral	0,13%
Desvio Padrão	1,82%
Valor de z	1,96
n	1259
Limite Superior	0,23%
Limite Inferior	0,03%
Excel	0,10%
Margem de Erro	0,10%

Hipótese nula	0,20%
P-valor	18,39%

Valor de z	=INV.T(1-0,05/2; C17-1)
n	1259



5. Limpeza de Dados

05 Limpeza de dados

Primeiro de tudo, esse treinamento é **mais focado no ferramental** necessário para que se possa **analisar dados**, incluindo principalmente **ferramentas de Excel avançado**, ignorando um pouco o Python pois este já foi passado em **outro treinamento** e tem o *script* explicado.

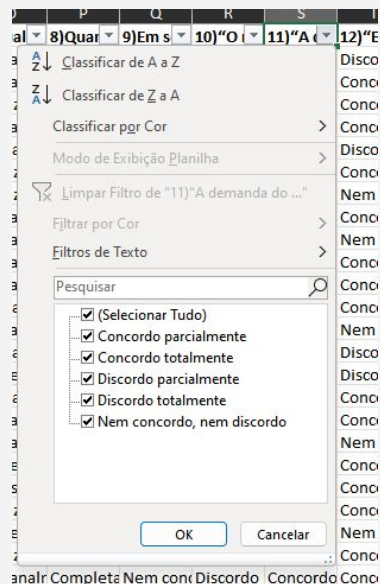
Alguns **comandos importantes** são necessários para **alterar dados**, são eles as **funções**: `se()`, `filtro()`, `seerro()`, `procv()`, `proch()`, `índice()`, `corresp()`, `média()`, `médiase()`, `somase()`, `cont.valores()`, `cont.se()`, `cont.ses()`, `esquerda()`, `direita()`, `arrumar()`.

Outras **ferramentas** também são abordadas para **alterar os dados no Excel**, que são: **filtro**, **tabela**, **tabela dinâmica**, **texto para colunas**, **preenchimento relâmpago (Ctrl + E)**, **remover duplicados**, **validação de dados** e **Power Query**. O Power Query será abordado em outro treinamento.



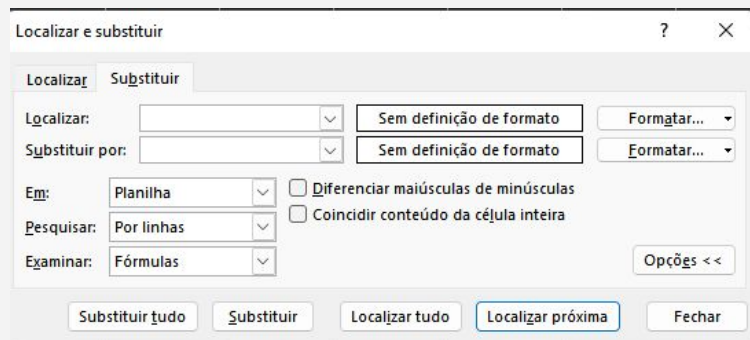
05 Limpeza de dados

Filtro (Ctrl + Shift + L): é uma das **funções mais importantes do Excel** e auxilia muito na análise de dados. Em uma tabela, normalmente pode-se **criar filtros** para todas as **variáveis** para alterar os dados dela, observando **quais os valores únicos** presentes e também podendo **ordenar os elementos na coluna**, além de filtrá-los.



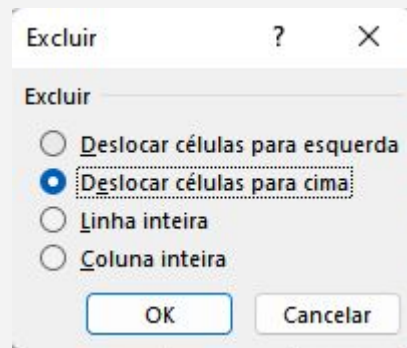
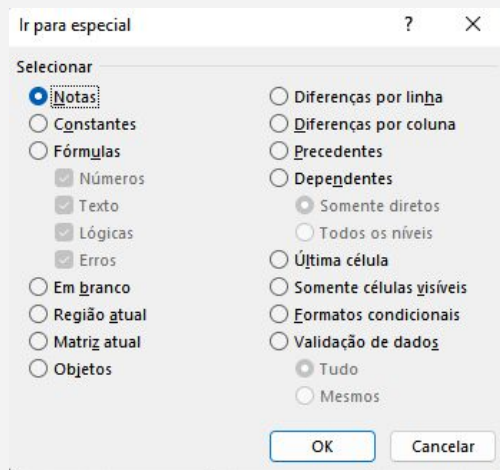
05 Limpeza de dados

Localizar e Substituir: esse comando é **bem instrutivo** e bem **importante** para a **limpeza de dados**, sua função é **substituir algum termo por outro** e pode ser usado em várias coisas, por exemplo **remover alguns espaços indesejados** da base (o código do Python de clusterização também faz isso automaticamente). É a função para substituir os **“Concordo totalmente” por 5** por exemplo.



05 Limpeza de dados

Ir para especial: tem **funções parecidas do que o localizar**, mas tem alguns objetivos diferentes. O **principal uso** dessa ferramenta é **remover células em branco** de uma tabela. Por exemplo, se quiser **remover células em branco**, vá para **Ir para especial** -> **Em branco** -> **Botão direito** -> **Excluir** -> **Deslocar células para cima**.



05 Limpeza de dados

=SE(teste_lógico; valor_se_verdadeiro; valor_se_falso): a **função SE** primeira testa **algum argumento lógico** e você diz o que **ele deve fazer caso o teste seja verdadeiro** e caso seja **falso**. Nesse sentido, ele é muito bom para diversas situações, até mesmo para **alterar as bases de dados**.

=FILTRO(matriz; incluir; [se_vazia]): essa função serve para **filtrar uma tabela** e, por isso, também **retorna uma tabela**. Primeiro, você **insere a tabela dentro de matriz** e dá a **condição** para a função na parte de incluir, por exemplo **=FILTRO(C1:D6;D1:D6 > 4)** filtra a tabela apenas com **valores acima de 4** para a variável na coluna D.

=SEERRO(valor; valor_se_erro): essa função **também é muito parecida com a função se**, porém você diz um valor para ela, normalmente uma **função** que pode dar erro e, se der erro, você também diz o que **ela deve substituir caso contrário**.



05 Limpeza de dados

=PROCV(valor_procurado; matriz_tabela; núm_índice_tabela): essa é uma das **funções mais conhecidas** e serve para **cruzar dados** entre tabelas. Primeiro, quando você tem uma tabela e **quer procurar algum valor específico** nela, você **diz primeiro o valor** que está correlacionado com aquele que você procura, depois a **matriz** que estão todos os valores e depois o **índice da variável procurada** (possível colocar **=CORRESP()** aqui).

=PROCH(valor_procurado; matriz_tabela; núm_índice_tabela): é muito parecido com o **procv**, mas é utilizada para **cruzar dados na horizontal**, em que o número índice é uma das linhas da tabela

=ÍNDICE(matriz; núm_linha; núm_coluna): a **função índice** retorna o **elemento de linha e coluna a serem definidos**. Ela funciona muito bem com o **corresp** que diz a posição de algum termo.



05 Limpeza de dados

=CORRESP(valor_procurado; matriz_procurada): é uma função que **casa muito com o procv e com o índice**, pois diz a **exata posição** de algum valor procurado dentro de uma matriz. Nesse sentido, pode utilizar o corresp dentro de **núm_índice_tabela** no **procv** para **automatizar a entrada** (quando o valor do índice muda) ou usar índice e corresp.

=ESQUERDA(texto; núm_caract): retorna os **n caracteres** à **esquerda** do texto. Serve muito para **alterar textos** para uma outra forma.

=DIREITA(texto; núm_caract): retorna os **n caracteres** à **direita** do texto. Também é uma função muito boa para **alterar textos**.

=ARRUMAR(texto): função muito importante, pois ela **retira todos os espaços em branco** de um texto (muitas vezes em pesquisa quanti estão cheios de espaços). Também podem ser removidos por **Localizar e Substituir**.



05 Limpeza de dados

=MÉDIA(matriz): retorna a **média de uma sequência de números**.

=MÉDIASE(range; critério): retorna a **média de uma distribuição** de números porém dá um **critério** para cada um dos números, por exemplo, **médiase(D1:D9; "> 5")** calcula a média de um intervalo **desconsiderando números menores que 5**

=SOMASE(intervalo; critério): **soma** os números porém com **algum critério** para cada um dos números, do mesmo modo que ocorre a **médiase**.

=CONT.VALORES(matriz): **conta quantos valores não vazios** estão na base de dados. Também há um **cont.se** que diz também o **critério** para contar os valores.

=SOMASES(intervalo_soma; intervalo_critério1; critério1; ...): é uma **soma com mais restrições para colocar**, primeiro colocando o **intervalo a ser somado e depois critérios**.



05 Limpeza de dados

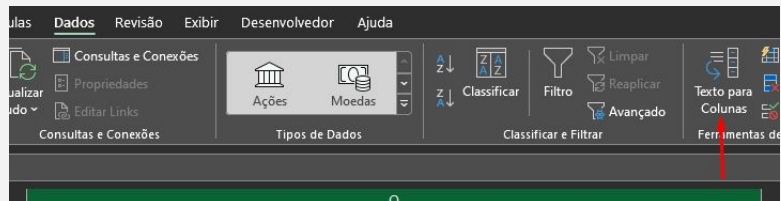
Preenchimento Relâmpago (Ctrl + E): é uma função muito importante no Excel que **ajuda na automatização de preenchimento de planilhas**. Nesse caso, você diz um padrão com base nos outros dados da tabela e aperta **Ctrl + E** ou vai em **Dados -> Preenchimento Relâmpago** que ele completa todos os outros dados. Isso é muito bom para **limpar dados** sem o uso de fórmulas.

A	B	C	D
Nome	Nascimento	Mês Nascimento	Ano Nascimento
Esthevão	03/ago	ago	2002
Enzo	12/set	set	2002
Manô	22/out		
Kike	01/dez		
Deco	10/jan		
Paraíba	19/fev		
Beto	31/mar		
Ana	10/mai		
Carol	19/jun		
Manu	29/jul		
Jorge	07/set		
Rafa	17/out		
Valen	26/nov		

A	B	C	D
Nome	Nascimento	Mês Nascimento	Ano Nascimento
Esthevão	03/ago	ago	2002
Enzo	12/set	set	2002
Manô	22/out	out	2002
Kike	01/dez	dez	2002
Deco	10/jan	jan	2002
Paraíba	19/fev	fev	2002
Beto	31/mar	mar	2002
Ana	10/mai	mai	2002
Carol	19/jun	jun	2002
Manu	29/jul	jul	2002
Jorge	07/set	set	2002
Rafa	17/out	out	2002
Valen	26/nov	nov	2002

05 Limpeza de dados

Texto para colunas: é uma ferramenta que separa uma coluna em várias bifurcando o texto presente. Vá em “Dados” -> “Texto para colunas” -> “delimitado” (para dizer algum separador que deve separar os textos) e seleciona qual esse delimitador (ponto e vírgula, espaço, barra são os mais comuns, também possível colocar o desejado em “outros”), dando **ok** depois (é importante não ter colunas do lado para ele conseguir separar).



4)Qual é o principal diferencial do seu estabelecimento?

Baixos preços; Qualidade dos produtos; Agilidade com o cliente
Qualidade dos produtos; Fornecimento de produtos premium ao consumidor; Ambiente agradável!
Pontualidade com entregas
Qualidade dos produtos; Agilidade com o cliente; Pontualidade com entregas
Baixos preços
Qualidade dos produtos; Pontualidade com entregas; Atendimento bom
Qualidade dos produtos; Atendimento bom; Forte presença da tecnologia
Baixos preços; Qualidade dos produtos; Agilidade com o cliente
Baixos preços
Baixos preços; Qualidade dos produtos; Agilidade com o cliente

Q	R	S	T
4)Qual é o principal diferencial do seu estabelecimento?			
Baixos preços	Qualidade dos p	Agilidade com o cliente	
Qualidade dos produtos	Fornecimento c	Ambiente agradável	
Pontualidade com entregas			
Qualidade dos produtos	Agilidade com c	Pontualidade com entrega	
Baixos preços			
Qualidade dos produtos	Pontualidade c	Atendimento bom	
Qualidade dos produtos	Atendimento b	Forte presença da tecnolo	
Baixos preços	Qualidade dos	Agilidade com o cliente	
Baixos preços	Qualidade dos	Fornecimento de produtos	
Baixos preços	Agilidade com c	Ambiente agradável	
Qualidade dos produtos	Ambiente agrac	Atendimento bom	
Baixos preços			
Baixos preços	Meu estabelecimento não possui diferenciais		
Baixos preços	Qualidade dos	Atendimento bom	
Baixos preços	Qualidade dos	Ambiente agradável	
Ambiente aeradável	Atendimento b	Meu estabelecimento não	

05 Limpeza de dados

Remover duplicadas: outra **ferramenta muito importante** para descobrir quais os elementos presentes em uma tabela. O **filtro** funciona muito bem nesses casos, mas muitas vezes é necessário **saber quais os elementos únicos nessa tabela**.

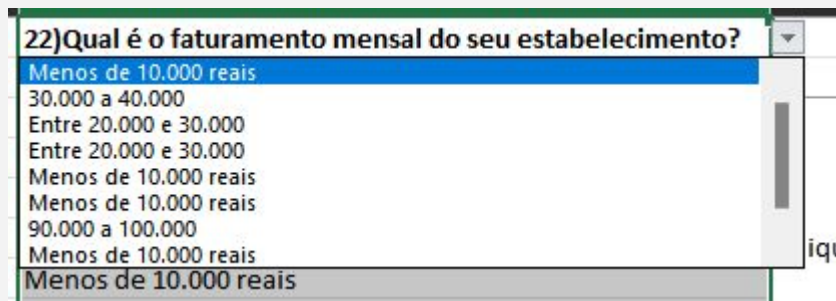
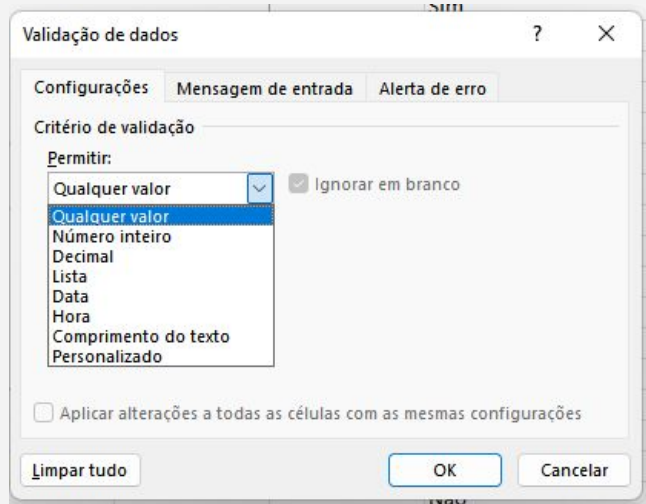
Depois de **removidas** as **duplicadas**, é possível **juntar isso com um cont.se** para saber **quantos valores têm em cada elemento**. Essa é uma das funções mais utilizadas, visto que para todas as **variáveis não clusterizáveis** é bom ter uma noção macro de distribuição.

Q	R	S	T
4)Qual é o principal diferencial do seu estabelecimento?			
Baixos preços	Qualidade dos p	Agilidade com o cliente	
Qualidade dos produtos	Fornecimento c	Ambiente agradável	
Pontualidade com entregas			
Qualidade dos produtos	Agilidade com c	Pontualidade com entrega	
Baixos preços			
Qualidade dos produtos	Pontualidade c	Atendimento bom	
Qualidade dos produtos	Atendimento b	Forte presença da tecnologia	
Baixos preços	Qualidade dos p	Agilidade com o cliente	
Baixos preços			
Baixos preços	Qualidade dos p	Agilidade com o cliente	
Baixos preços	Qualidade dos p	Fornecimento de produtos	
Qualidade dos produtos	Agilidade com c	Ambiente agradável	
Baixos preços	Ambiente agrac	Atendimento bom	
Baixos preços			
Meu estabelecimento não possui diferenciais			
Baixos preços	Qualidade dos p	Atendimento bom	
Baixos preços	Qualidade dos p	Ambiente agradável	
Ambiente agradável	Atendimento b	Meu estabelecimento não	

Q
4)Qual é o principal diferencial do seu estabelecimento?
Baixos preços
Qualidade dos produtos
Pontualidade com entregas
Meu estabelecimento não possui diferenciais
Ambiente agradável
Fornecimento de produtos premium ao consumidor
Atendimento bom
Agilidade com o cliente

05 Limpeza de dados

Validação de dados: essa é uma ferramenta de **formatação das células**. Normalmente, essa **validação de dados** normalmente é utilizada para **transformar células** em listas que serão preenchidas pelo cliente depois. É muito semelhante com a **caixa de seleção do VBA**, porém no próprio Excel.



05 Limpeza de dados

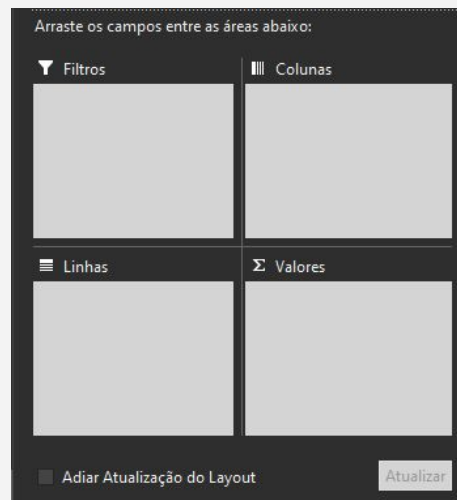
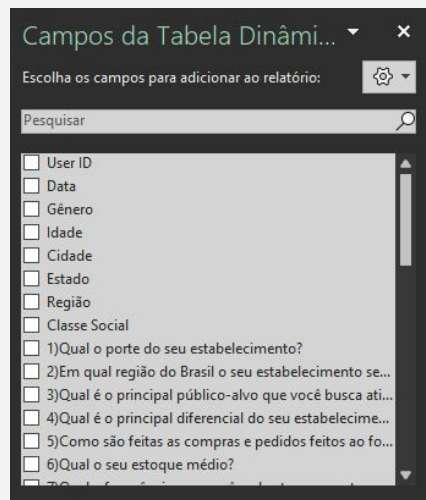
Formatar como tabela: é algo **muito importante no Excel**, pois facilita a **manipulação de dados**. Depois de **formatar como tabela**, é possível, por exemplo, **adicionar novas linhas** na tabela mantendo a formatação da tabela como um todo. Outro fator importante é que a **tabela puxa também a fórmula usada nas outras linhas** também para a nova linha da tabela, como mostrado abaixo. Ela permite criar **gráficos automáticos** que mudam quando mudar o filtro de alguma coluna.

Nome	=ESQUERDA("A1"; 3)
Esthevão	Est
Enzo	Enz
Manô	Man
Kike	Kik
Deco	Dec
Paraíba	Par
Beto	Bet
Ana	Ana
Carol	Car
Manu	Man
Jorge	Jor
Rafa	Raf
Valen	Val

Nome	=ESQUERDA("A1"; 3)
Esthevão	Est
Enzo	Enz
Manô	Man
Kike	Kik
Deco	Dec
Paraíba	Par
Beto	Bet
Ana	Ana
Carol	Car
Manu	Man
Jorge	Jor
Rafa	Raf
Valen	Val
Eduardo	Edu

05 Limpeza de dados

Tabela dinâmica: também muito importante para **analisar dados**, foi explicada no **treinamento de dashboard** mas ela também tem muitas funções para a análise em **pesquisas quantitativas**. Após formatar como tabela, vá em **Design da Tabela -> Resumir em Tabela Dinâmica -> Planilha Existente** e escolher uma célula em uma outra aba.



05 Limpeza de dados

Tabela dinâmica: a **tabela** tem dois principais campos, um com todas as **variáveis** e outro com **4 campos para criar a tabela**. Normalmente adiciona-se variáveis em **linhas, colunas e valores** para poder se analisar como estão distribuídas, além de **criar relações** entre variáveis. Ela faz o processo de **remover duplicatas e cont.se** de um jeito automático mas nem sempre é possível fazê-la.

Campos da Tabela Dinâmi... ✕

Escolha os campos para adicionar ao relatório: ⚙

Pesquisar 🔍

- ☐ User ID
- ☐ Data
- ☐ Gênero
- ☐ Idade
- ☐ Cidade
- ☐ Estado
- ☐ Região
- ☐ Classe Social
- ☐ 1)Qual o porte do seu estabelecimento?
- ☐ 2)Em qual região do Brasil o seu estabelecimento se...
- ☐ 3)Qual é o principal público-alvo que você busca ati...
- ☐ 4)Qual é o principal diferencial do seu estabelecime...
- ☐ 5)Como são feitas as compras e pedidos feitos ao fo...
- ☐ 6)Qual o seu estoque médio?

Arraste os campos entre as áreas abaixo:

▼ Filtros

▮ Colunas

▮ Linhas

Σ Valores

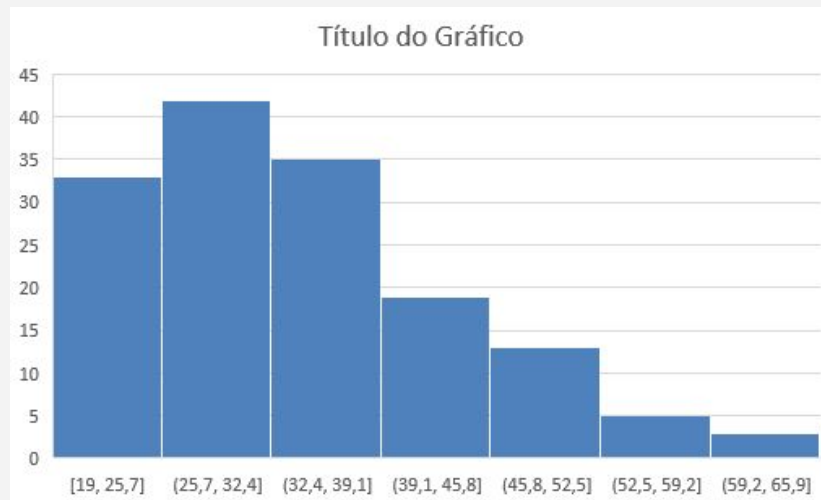
Adiar Atualização do Layout

Atualizar

Contagem de Gênero		Rótulos de Coluna	
Rótulos de Linha	Feminino	Masculino	Total Geral
Centro Oeste	8	5	13
Nordeste	20	22	42
Norte	7	5	12
Sudeste	44	27	71
Sul	8	4	12
Total Geral	87	63	150

05 Limpeza de dados

Histograma: para **análise de dados numéricos**, é bom fazer um **histograma** para entender o **comportamento** deles. O histograma funciona como um **complemento** para os **boxplots** na análise de como estão **compartimentados**. No exemplo abaixo, é possível perceber que as **pessoas estão mais presentes em anos mais novos**.





Obrigado!