Guia Excel: Limpeza e Análise de Dados

Existem duas etapas principais para realizar uma análise: limpeza de dados e análise de dados. Este documento guiará você por essas etapas usando o Microsoft Excel.[1] Neste exercício, vamos limpar e analisar os resultados de uma avaliação hipotética para ver como uma versão modificada de um e-mail impacta o número de destinatários que clicam em um link no e-mail. Nesta simulação, os indivíduos do grupo de tratamento receberam o e-mail modificado e os indivíduos do grupo de controle receberam o e-mail normal.

Pré-limpeza

Antes de trabalhar com um conjunto de dados, você deve criar uma cópia dos dados brutos e revisar os dados para entender quais informações o conjunto contém e quais alterações podem ser necessárias.

Criando uma cópia dos dados brutos
Ao trabalhar com dados, é uma prática recomendada preservar todas as informações em um arquivo de dados não editado. Antes de executar qualquer limpeza ou análise de dados, salve os dados brutos (ou seja, os dados originais não editados) e faça uma cópia do arquivo para você editar. Isso evita a perda de dados que podemos precisar para análises posteriores ou para verificar se não cometemos erros (e.g., substituir involuntariamente um valor de idade de 45 por 2).

Revisando dados
As planilhas geralmente contêm muitas colunas de informações e podem incluir códigos ou abreviações (e.g., “Feminino” pode ser codificado como “F”, 0 ou 1), e os campos podem não estar formatados corretamente. Antes de limpar seu conjunto de dados, verifique seus dados para entender quais informações o conjunto de dados contém e identificar o que você pode precisar limpar.

Ao revisar os dados, faça a si mesmo as seguintes perguntas:

- Quantas observações totais (linhas) existem no meu conjunto de dados?
- O que cada observação (linha) representa (e.g., uma pessoa, uma família, uma escola)?
- Existem colunas para todos os dados que preciso/esperava ter para a análise?
- Quais colunas eu preciso para esta análise? Quais colunas eu não preciso e posso excluir?
- Eu entendo qual variável cada coluna representa?
    - Ex: med_inc = Renda média
- Eu sei o que significam os valores em cada coluna?

[1] Este guia foi criado usando o Microsoft Excel 2021 Versão 16.57 e os comandos de teclado são definidos para o Windows. Os locais das funções do Excel podem variar dependendo da versão do Excel que você está usando, e os comandos de teclado serão diferentes para outros sistemas operacionais.

    - *Ex:* Para uma coluna de raça com valores codificados como 1, 2, 3, 4, etc., a qual raça cada valor corresponde?
- Os valores são formatados consistentemente em cada coluna?
    - *Ex:* Os nomes estão todos em maiúsculas, minúsculas, formato apropriado ou uma mistura?
- Os valores estão formatados de maneira conveniente para análise? Se não, como eles devem ser formatados? *(Consulte o Conceito 1 em Limpeza de dados para ver um exemplo.)*
- Existem observações duplicadas? *(Consulte o Conceito 5 em Limpeza de dados.)*
- Existem valores ausentes para colunas importantes? *(Consulte o Conceito 4 em Limpeza de dados.)*

Limpeza de dados

Depois de revisar seus dados, você pode excluir colunas desnecessárias para a análise. No restante deste documento, revisaremos alguns conceitos para executar a limpeza de dados no Excel.

O que você vai aprender a fazer?

1. Criar variáveis numéricas a partir de variáveis de texto
2. Padronizar a formatação
3. Remover/substituir caracteres
4. Verificar se há valores ausentes
5. Verificar se há observações duplicadas
6. Remover observações duplicadas

Conceito 1. Criar variáveis numéricas a partir de variáveis de texto

É importante confirmar que o status de tratamento e as variáveis de resultado sejam indicados como zeros e uns. Por exemplo, se você está medindo se os destinatários de e-mail clicaram em um link, aqueles que clicaram no link devem ser indicados como “1”, enquanto aqueles que não clicaram devem ser indicados como “0” no Excel.

Se o seu status de tratamento ou variável de resultado ainda não estiver registrado em zeros e uns (por exemplo, é comum que o resultado seja registrado como “sim” ou “não”), o Excel pode convertê-los rapidamente em zeros e uns com o seguinte comando::

=SE(E2=”sim”;1;0)

Passo 1a. Ao lado da coluna denominada “resultado”, nomeie esta coluna como “indicador de resultado” e, na célula F2, insira o comando acima.

Imagem 1. Atribuindo um valor para indicador de resultado

Email	rand1	rand2	tratamento	resultado	indicador de resultado
klm021@gmail.com	0.1067483	0.8221911	0	yes	=SE(E2=”sim”;1;0)
hij789@hotmail.com	0.5016783	0.9384599	0	no	
abd123@gmail.com	0.2305955	0.2684939	1	yes	
efg456@yahoo.com	0.7142028	0.4684949	1	yes	

Este comando faz com que o Excel realize um teste lógico: se o valor de E2 (coluna E é o campo que registra o resultado neste exemplo) for “sim”, então indique “1”, caso contrário, indique “0”.

Passo 1b. Agora, preencha este comando em toda a coluna para converter todos os resultados em zeros e uns. Você pode fazer isso manualmente ou usando comandos de teclado.

Preenchimento manual: selecione a célula onde você inseriu o comando (neste caso, E2) e arraste a alça de preenchimento  (o ponto que aparece no canto inferior direito) até a linha final.

Imagem 2. Preenchimento manual de um comando em uma coluna

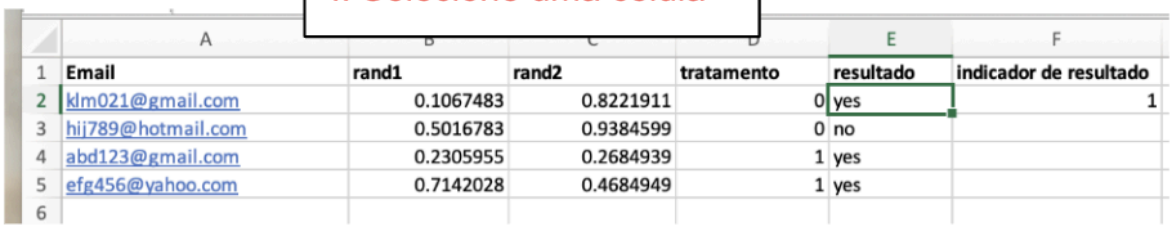
Email	rand1	rand2	tratamento	resultado	indicador de resultado
klm021@gmail.com	0.1067483	0.8221911	0	yes	1
hij789@hotmail.com	0.5016783	0.9384599	0	no	0
abd123@gmail.com	0.2305955	0.2684939	1	yes	1
efg456@yahoo.com	0.7142028	0.4684949	1	yes	1

Preenchimento usando comandos do teclado. Se o conjunto de dados for tão grande que seja difícil rolar manualmente até o final, você também pode preencher este comando em toda a coluna usando os comandos do teclado. (Ver imagem 3 abaixo.)

i. Selecione uma célula em uma coluna preenchida, de preferência a coluna ao lado da que deseja preencher (neste caso, “resultado”).
ii. Pressione Ctrl + Seta para baixo. Isso leva você à linha final do conjunto de dados.
iii. Selecione a célula à direita da última célula da coluna “resultado”. Esta será a célula final da coluna “indicador de resultado”.
iv. Pressione Shift + Ctrl + Seta para cima. Isso seleciona da célula final até a célula preenchida mais próxima. Neste caso, E2 que contém nosso comando.
v. Preencha o comando pressionando Ctrl + D.

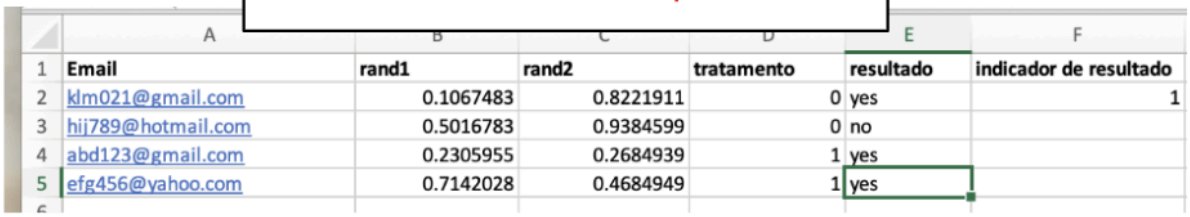
Imagem 3. Preenchendo um comando em uma coluna usando comandos do teclado

I. Selecione uma célula



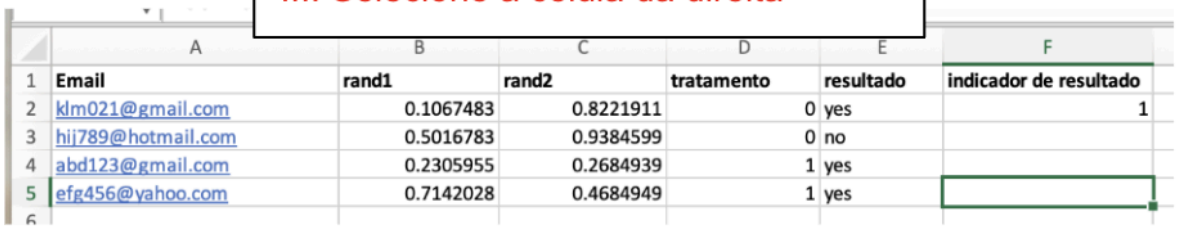
	A	B	C	D	E	F
1	**Email**	**rand1**	**rand2**	**tratamento**	**resultado**	**indicador de resultado**
2	klm021@gmail.com	0.1067483	0.8221911	0	yes	1
3	hij789@hotmail.com	0.5016783	0.9384599	0	no	
4	abd123@gmail.com	0.2305955	0.2684939	1	yes	
5	efg456@yahoo.com	0.7142028	0.4684949	1	yes	
6						

II. Pressione Ctrl + Seta para baixo



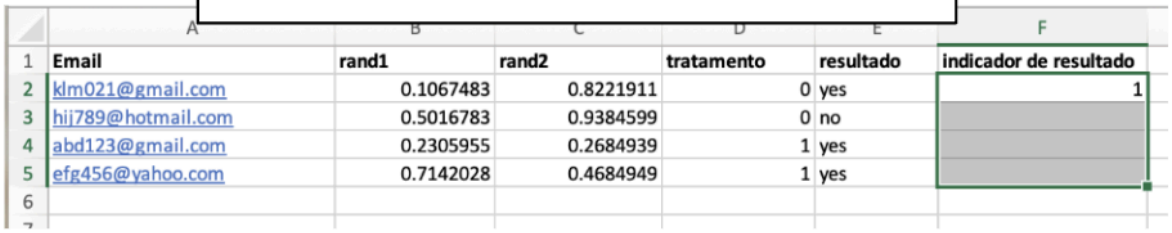
	A	B	C	D	E	F
1	**Email**	**rand1**	**rand2**	**tratamento**	**resultado**	**indicador de resultado**
2	klm021@gmail.com	0.1067483	0.8221911	0	yes	1
3	hij789@hotmail.com	0.5016783	0.9384599	0	no	
4	abd123@gmail.com	0.2305955	0.2684939	1	yes	
5	efg456@yahoo.com	0.7142028	0.4684949	1	yes	

III. Selecione a célula da direita



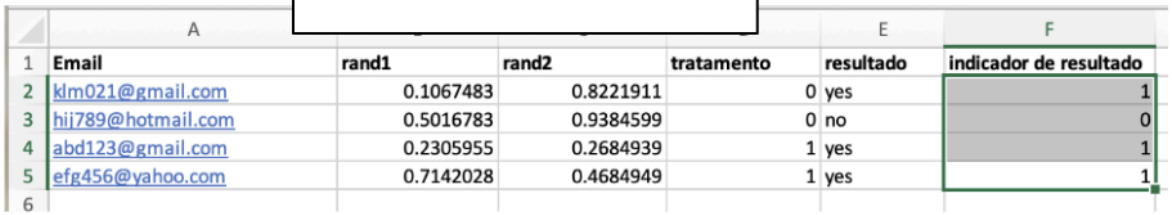
	A	B	C	D	E	F
1	**Email**	**rand1**	**rand2**	**tratamento**	**resultado**	**indicador de resultado**
2	klm021@gmail.com	0.1067483	0.8221911	0	yes	1
3	hij789@hotmail.com	0.5016783	0.9384599	0	no	
4	abd123@gmail.com	0.2305955	0.2684939	1	yes	
5	efg456@yahoo.com	0.7142028	0.4684949	1	yes	

IV. Pressione Shift + Ctrl + Seta para cima



	A	B	C	D	E	F
1	**Email**	**rand1**	**rand2**	**tratamento**	**resultado**	**indicador de resultado**
2	klm021@gmail.com	0.1067483	0.8221911	0	yes	1
3	hij789@hotmail.com	0.5016783	0.9384599	0	no	
4	abd123@gmail.com	0.2305955	0.2684939	1	yes	
5	efg456@yahoo.com	0.7142028	0.4684949	1	yes	
6						

V. Pressione Ctrl + D



	A	B	C	D	E	F
1	**Email**	**rand1**	**rand2**	**tratamento**	**resultado**	**indicador de resultado**
2	klm021@gmail.com	0.1067483	0.8221911	0	yes	1
3	hij789@hotmail.com	0.5016783	0.9384599	0	no	0
4	abd123@gmail.com	0.2305955	0.2684939	1	yes	1
5	efg456@yahoo.com	0.7142028	0.4684949	1	yes	1
6						

Depois de preencher o comando, temos uma nova coluna de zeros e uns que o Excel pode usar prontamente para análise.

Conceito 2. Padronizar a formatação

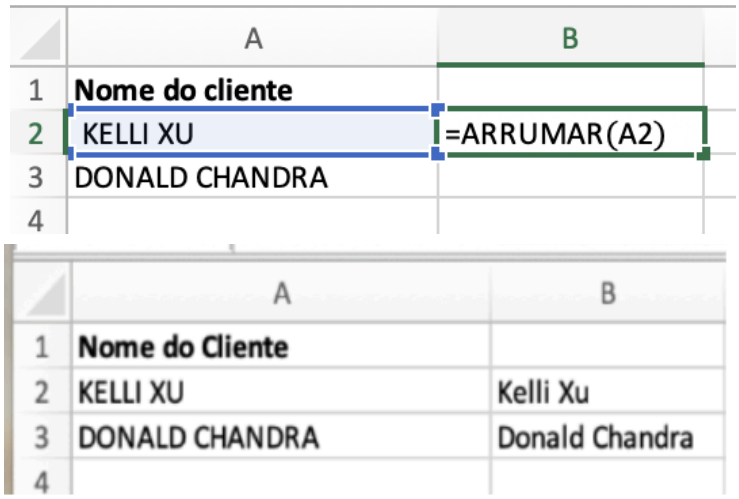
Isso é especialmente importante para nomes próprios. Posteriormente, aprenderemos como verificar e remover entradas duplicadas. É importante padronizar a formatação primeiro, antes de remover duplicatas. Para entender o porquê, imagine que existam entradas duplicadas para uma pessoa chamada Kelli Xu, mas a formatação é diferente (por exemplo, “Kelli Xu” vs. “KELLI XU”). O Excel não irá tratá-las como valores duplicados até que você as padronize.

Altere a aparência do texto usando os seguintes comandos:

- =MAIÚSCULA() - converte todas as letras do texto em maiúsculas
- =MINÚSCULA() - converte todas as letras do texto em minúsculas
- =PRI.MAIÚSCULA() - converte o texto para que a primeira letra de cada palavra seja maiúscula; o resto, minúscula

Nota: para nomes, queremos usar letras maiúsculas “adequadas” (use a função “PRI.MAIÚSCULA()”). Preencha este comando na coluna usando a técnica descrita acima para atribuir valores de indicador de resultado.

Imagem 4. Alterando valores maiúsculos para formatação adequada



Conceito 3. Remover/substituir caracteres

Passo 3a. Remova espaços extras e caracteres não imprimíveis. Esta etapa também é importante antes de remover duplicatas. Imagine se o valor duplicado para Kelli Xu contiver um espaço incorreto no final do nome (por exemplo, “Kelli Xu” vs. “Kelli Xu ”), o Excel não os tratará como valores duplicados.

Passo 3b. Substitua os caracteres com acentos por caracteres sem acento. Por exemplo, o Excel não tratará “Maria Ramirez” como um valor duplicado de “Maria Ramírez”.

Identifique, remova e substitua caracteres usando os seguintes comandos/funções:

- Use a função “ARRUMAR()” para remover espaços irrelevantes do texto.
- Use “**Localizar e substituir**” para identificar caracteres específicos e substituí-los por outros caracteres (por exemplo, encontre “ñ” e substitua por “n”).

Conceito 4. Verificar se há valores ausentes

Frequentemente, há valores ausentes nos conjuntos de dados. Às vezes, esses valores estão ausentes para variáveis menos importantes da análise. Por exemplo, não há problema se não tivermos os nomes do meio de alguns dos destinatários de nossos e-mails, porque isso não afeta nossa compreensão do efeito dos e-mails nos destinatários que clicam em um link. Por outro lado, se não tivermos certeza de qual versão de nosso e-mail um grupo de destinatários recebeu porque faltam dados de atribuição de tratamento, isso dificulta a identificação da versão mais eficaz.

Uma boa maneira de verificar se há valores ausentes é usar a formatação condicional para destacá-los.

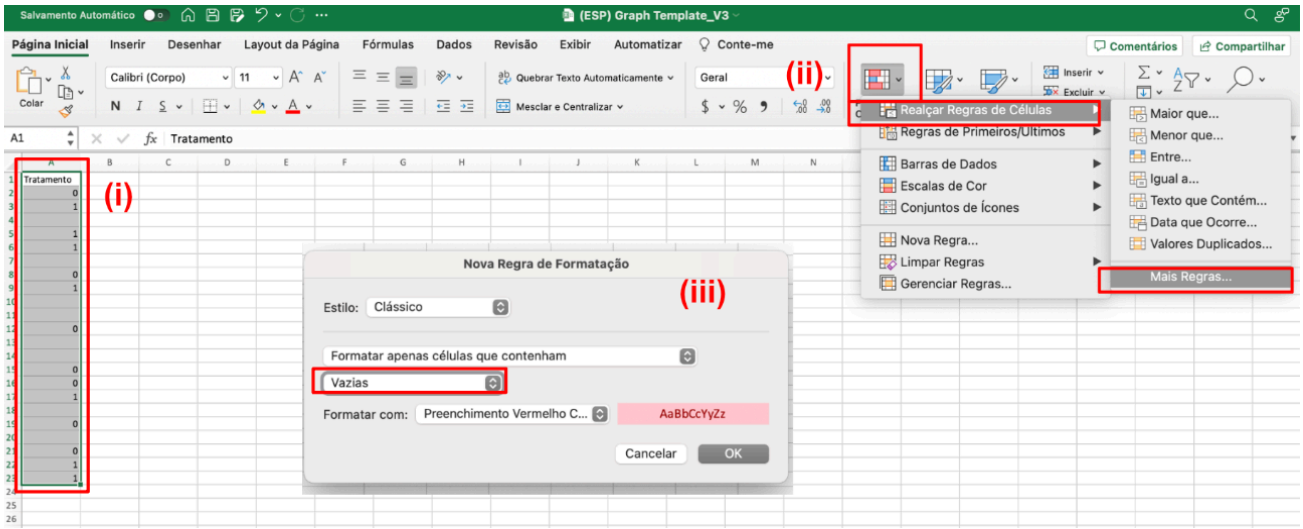
Para destacar os valores ausentes (Ver imagem 5.)

i. Selecione o intervalo de valores que você gostaria de verificar. Isso pode ser para uma variável, múltiplas variáveis ou todo o conjunto de dados. *Não selecione a(s) coluna(s) inteira(s), caso contrário, todas as células vazias abaixo do seu conjunto de dados serão realçadas.*

ii. Na guia Página Inicial, selecione Formatação Condicional >> Regras de Realce das Célula >> Mais Regras…

iii. Na caixa abaixo de “Formatar apenas as células com”, selecione “Vazias” e clique em OK. Todas as células vazias devem ser realçadas.

Imagem 5. Formatação condicional para valores ausentes



	A	B	C	D
1	Email	rand1	rand2	tratam
2	qbu397@gmail.com	0.03800071	0.04954939	0
3	ipb258@hotmail.com	0.7766173	0.14906559	1
4	ung938@hotmail.com	0.59828285	0.15296501	
5	dfc273@hotmail.com	0.74523049	0.21603585	1
6	fkr920@yahoo.com	0.89660444	0.26220002	1
7	ooz622@yahoo.com	0.22615665	0.2967678	
8	ynn124@gmail.com	0.60745339	0.3287076	
9	efg456@yahoo.com	0.08847096	0.37850433	0
10	kfg443@gmail.com	0.78139934	0.48369506	1
11	ysz508@gmail.com	0.16307376	0.48681567	
12	ghk743@gmail.com	0.60307674	0.49502387	
13	qsk252@gmail.com	0.01290083	0.54788977	0
14	qff249@gmail.com	0.62388468	0.55775251	
15	abd123@gmail.com	0.67501482	0.56021215	
16	xil729@gmail.com	0.09315082	0.56222449	0
17	cel175@gmail.com	0.00383966	0.5663684	0
18	jlm744@gmail.com	0.98356611	0.6457943	1
19	klm021@gmail.com	0.20965767	0.65050048	
20	ski867@hotmail.com	0.091135	0.75130712	0
21	oml387@gmail.com	0.2188718	0.7570109	
22	jgi224@gmail.com	0.07370952	0.84975785	0
23	hqh661@gmail.com	0.81677	0.863426	1
24	hdm621@gmail.com	0.76886625	0.90409789	1

Para classificar valores ausentes (i.e. trazer observações com valores ausentes para o topo do seu conjunto de dados)

i. Selecione todos os valores em seu conjunto de dados.
ii. Na guia Página Inicial, selecione Classificar e Filtrar >> Filtro.
iii. Clique na seta para baixo no rótulo da coluna de interesse.
iv. Na opção “Classificar por cor”, escolha o formato de célula realçado.

Para filtrar valores ausentes (i.e. visualizar apenas observações com valores ausentes)
Nota: quando você filtra as observações, elas são ocultadas, não excluídas.

i. Selecione todos os valores em seu conjunto de dados.
ii. Na guia Página Inicial, selecione Classificar e Filtrar >> Filtro.
iii. Clique na seta para baixo no rótulo da coluna de interesse.
iv. No menu “Filtrar por cor”, escolha o formato de célula realçado. Você verá apenas observações com valores ausentes nesta coluna.

Nota: Para limpar a formatação condicional, na guia Página Inicial, selecione Formatação Condicional >> Limpar Regras >> Limpar Regras da Planilha Inteira ou Limpar Regras das Células Selecionadas.

Conceito 5. Verificar se há observações duplicadas

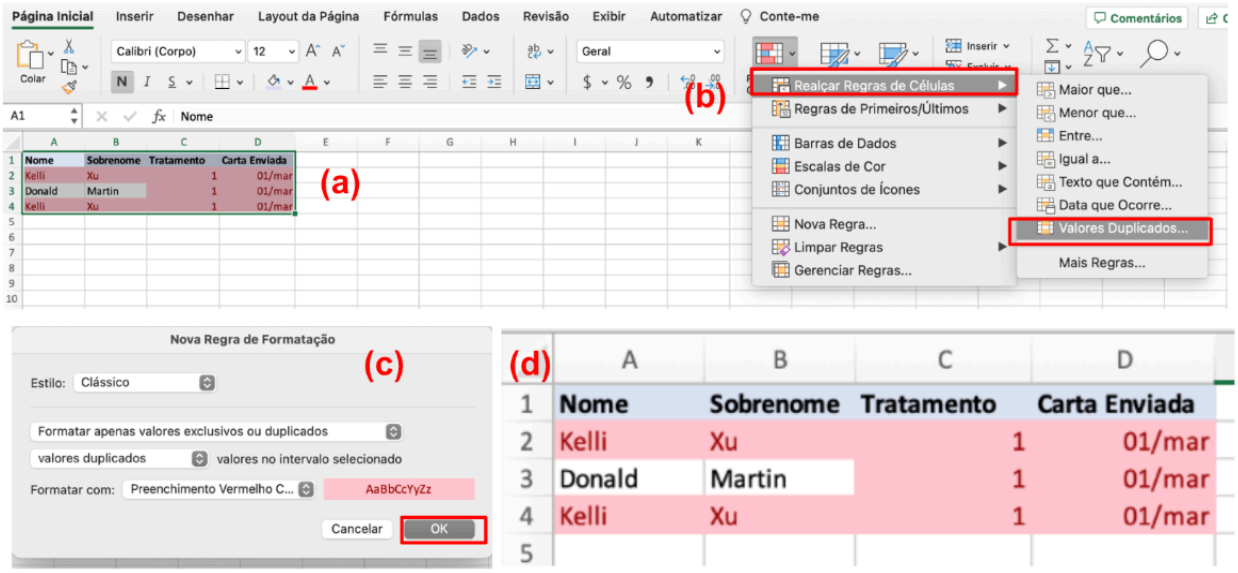
Verifique nomes e endereços para ver se há duplicatas. Se houver linhas duplicadas completamente idênticas, exclua-as.

Vamos pegar, por exemplo, um conjunto de dados onde cada observação é uma pessoa, e temos seu nome completo. Haverá muitos primeiros nomes duplicados e pode haver sobrenomes duplicados. Geralmente queremos identificar se o *nome completo* é ou não uma

duplicata. Se o conjunto de dados já incluir uma coluna com o nome completo de uma pessoa, escolha essa variável para localizar ou remover duplicatas. Se uma coluna incluir o primeiro nome de uma pessoa e outra coluna incluir seu sobrenome, escolha ambas as colunas ao verificar se há duplicatas. (Para os passos abaixo, consulte a imagem 6.)

Passo 5a. Selecione seus dados.
Passo 5b. Na guia Página Inicial, selecione Formatação Condicional >> Regras de Realce das Células >> Valores Duplicados…
Passo 5c. Na janela Nova Regra de Formatação, clique em OK. (Você pode alterar a cor do realce no menu suspenso “Formatar com”.
Passo 5d. Os valores duplicados agora serão realçados.

Imagem 6. Verificando se há valores duplicados



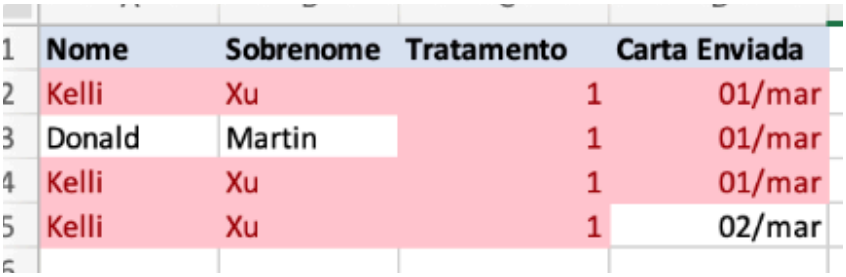
(!) No Conceito 6, você aprenderá a remover observações duplicadas, mas vamos refletir se as observações duplicadas fazem sentido ou se devemos excluí-las.

No exemplo acima, vemos que existem observações idênticas para Kelli Xu. Em um cenário em que nossa intervenção envolve o envio de uma carta por pessoa, poderíamos concluir que devemos excluir a observação duplicada para Kelli Xu. Não gostaríamos de contar seus resultados duas vezes ao analisar os dados.

No entanto, vamos imaginar uma intervenção na qual um grupo de participantes recebe uma carta e o outro grupo recebe a mesma carta E uma segunda carta de lembrete. Depois de destacar os valores duplicados, nosso conjunto de dados pode se parecer com a Figura 7 abaixo.

Gostaríamos de excluir a linha 4, que é uma duplicata exata da linha 2, mas não desejamos excluir a 3ª observação de Kelli Xu na linha 5, pois esta linha mostra quando sua segunda carta foi enviada.

Imagem 7. Verificando se há valores duplicados, continuação



	Nome	Sobrenome	Tratamento	Carta Enviada
2	Kelli	Xu	1	01/mar
3	Donald	Martin	1	01/mar
4	Kelli	Xu	1	01/mar
5	Kelli	Xu	1	02/mar

A chave é pensar sobre sua intervenção e depois se perguntar: “Faz sentido que existam valores duplicados?”

Conceito 6. Remover observações duplicadas

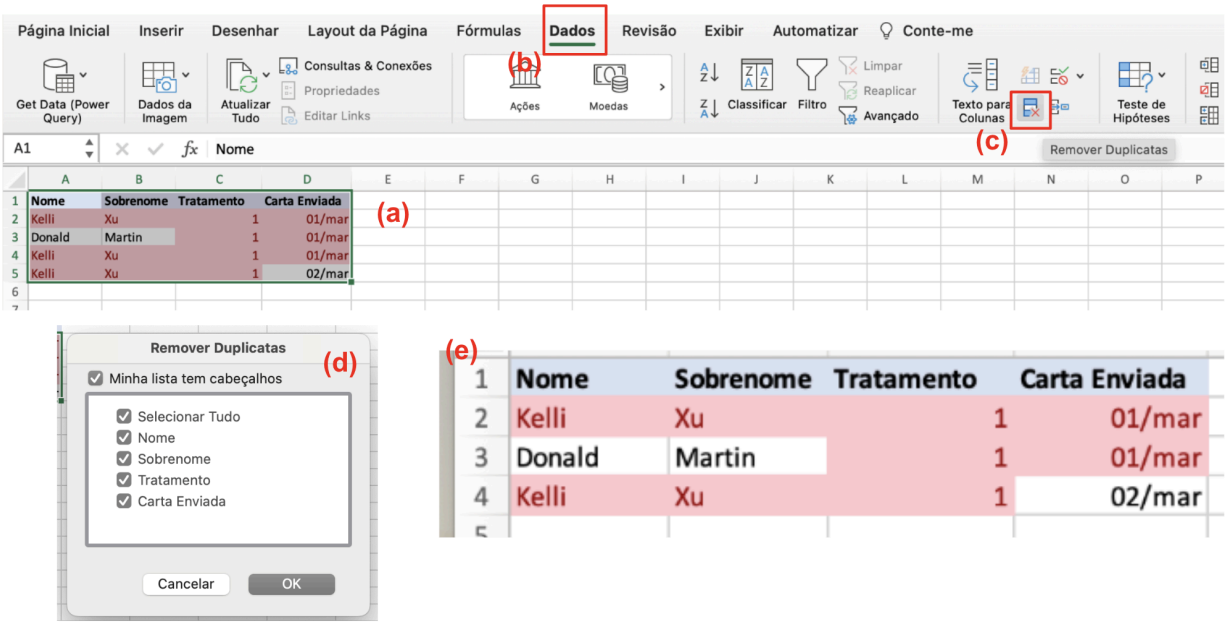
Se você acha que uma observação duplicada deve ser excluída, siga as etapas abaixo para removê-la. (Ver imagem 8.)

Passo 6a. Selecione seus dados.
Passo 6b. Vá para a guia Dados
Passo 6c. Clique no botão Remover Duplicatas.
Passo 6d. Na janela Remover Duplicatas, verifique se a opção “Minha lista tem cabeçalhos” está marcada, se você selecionou todos os dados, incluindo os cabeçalhos. Em seguida, verifique se todas as colunas estão selecionadas. Clique em OK.
Passo 6e. As observações duplicadas serão excluídas..

(!) Observe que apenas a entrada duplicada para Kelli Xu recebendo a carta em 1º de março foi removida, mas não a entrada indicando que Kelli Xu recebeu a carta em 1º de abril. Isso ocorre porque selecionamos todas as colunas na janela Remover Duplicatas. Selecionar todas as colunas significa que a observação teria que ser uma duplicata em todas essas colunas.

Se não tivéssemos selecionado a Coluna D, o Excel teria verificado observações idênticas para as Colunas A, B e C. Nesse caso, as linhas 4 e 5 teriam sido removidas.

Imagem 8. Removendo entradas duplicadas



Algumas outras funções que podem ser úteis.
Strings são simplesmente partes de texto que podem ser manipuladas, consultadas, movidas e editadas usando funções padrão adicionais do Excel. Como exemplo dessas funções, faremos referência aos dados da imagem abaixo.

	A	B
1	Nome	Sobrenome
2	Kelli	Xu

- CONCATENAR() – agrupa várias strings
  - Ex: =CONCATENAR(A2;" ";B2) gera "Kelli Xu".
- ESQUERDA() – retorna os nº de caracteres à esquerda de uma string
  - Ex: =ESQUERDA(A2;3) gera "Kel".
- DIREITA() – retorna os nº de caracteres à direita de uma string
  - Ex: =DIREITA(A2;2) gera "li".
- EXT.TEXTO() – retorna caracteres do meio de uma string
  - Ex: =EXT.TEXTO(A2;2;3) gera "ell".

Análise de dados

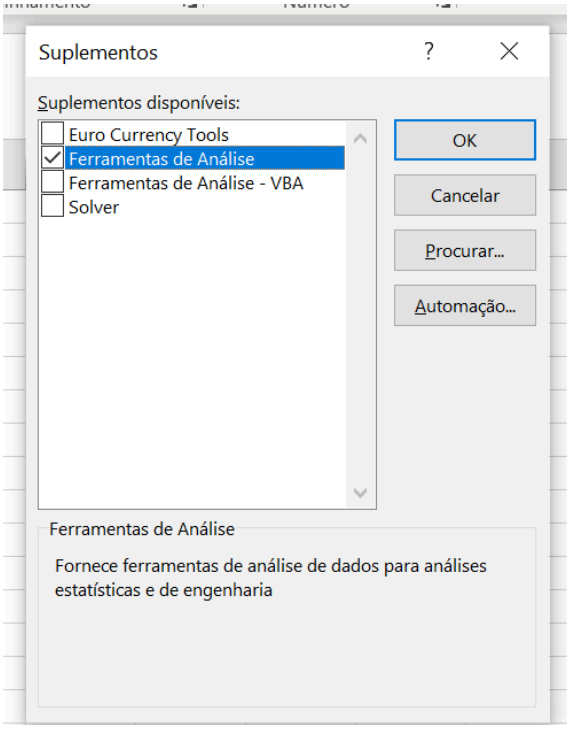
Depois de limpar os dados, você está pronto para começar a analisá-los. A análise de dados nos ajuda a entender melhor nossos dados, descobrir padrões, tirar conclusões e fundamentar a tomada de decisões. Esta seção fornece uma visão geral básica de como descrever dados e tirar conclusões de uma análise de regressão básica de um ensaio controlado randomizado.

O que você vai aprender a fazer?

1. Produzir e interpretar estatísticas descritivas
2. Executar uma regressão
3. Interpretar os resultados da regressão
4. Fazer um gráfico com os resultados da regressão

Para realizar a análise estatística, usaremos o Add-in Analysis Toolpak do Excel (Instruções de instalação).

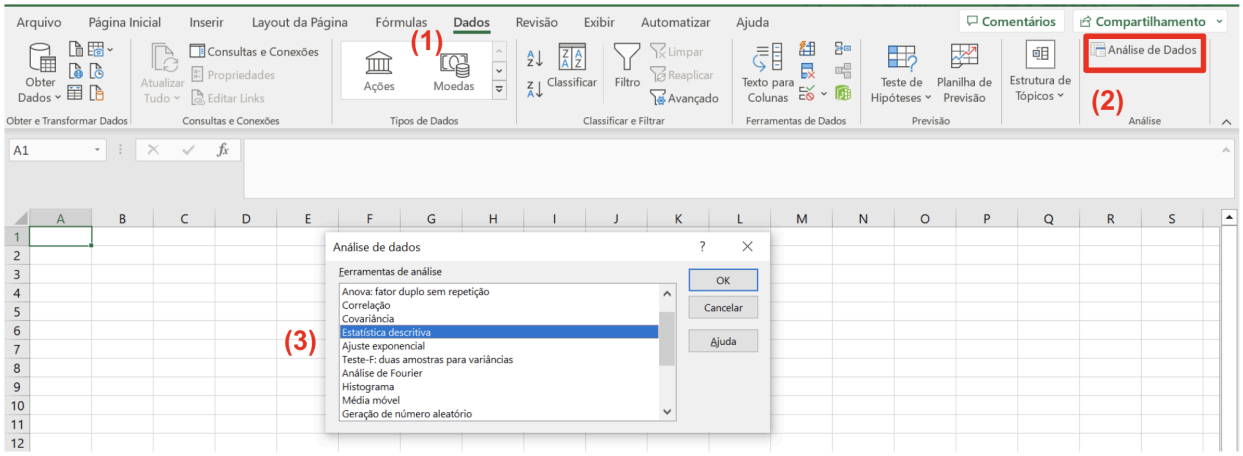
Imagem 9. Add-in Analysis Toolpak do Excel



Passo 1. Produzir e interpretar estatísticas descritivas

Após a instalação, na guia “Dados” (1), utilize a função “Análise de Dados” (2), e selecione “Estatística descritiva” (3). Clique em "OK". (Ver imagem 10.)

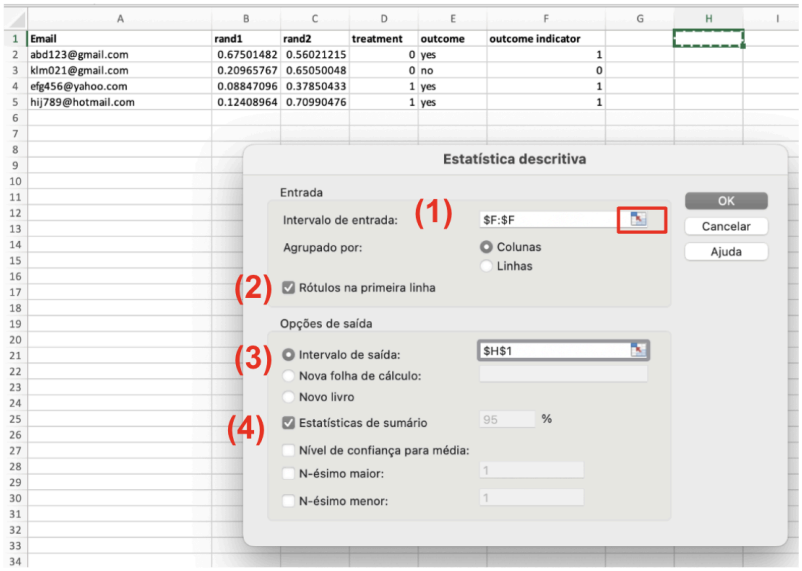
Imagem 10. Acessando a ferramenta Estatística Descritiva



Na Janela “Estatística descritiva”, selecione os dados da variável para a qual deseja produzir estatísticas descritivas como “Intervalo de entrada” clicando no botão  (1). Você pode a) escolher a coluna inteira (neste caso, clicando na coluna F) e marcar “Rótulos na primeira linha” (2) ou b) selecionar apenas os valores sob o rótulo da variável e certificar-se de que “Rótulos na primeira linha ” está desmarcado.

Em seguida, escolha onde deseja que a tabela de resumo estatístico apareça, clicando no botão  ao lado de “Intervalo de saída” (3). Você pode escolher onde deseja que a tabela apareça, e a célula selecionada será o canto superior esquerdo da tabela. Neste exemplo, selecionamos H1 como o início da tabela. Por fim, marque a opção “Estatísticas de sumário” (4) e clique em “OK”. (Ver imagem 11.)

Imagem 11. Produzindo as estatísticas descritivas



Depois de clicar em “OK”, a tabela de resumo estatístico para esta variável aparecerá onde você definiu o intervalo de saída e será semelhante às duas primeiras colunas da tabela na imagem 12. Na terceira coluna, incluímos descrições de conceitos comuns.

Imagem 12. Tabela de resumo estatístico com descrições

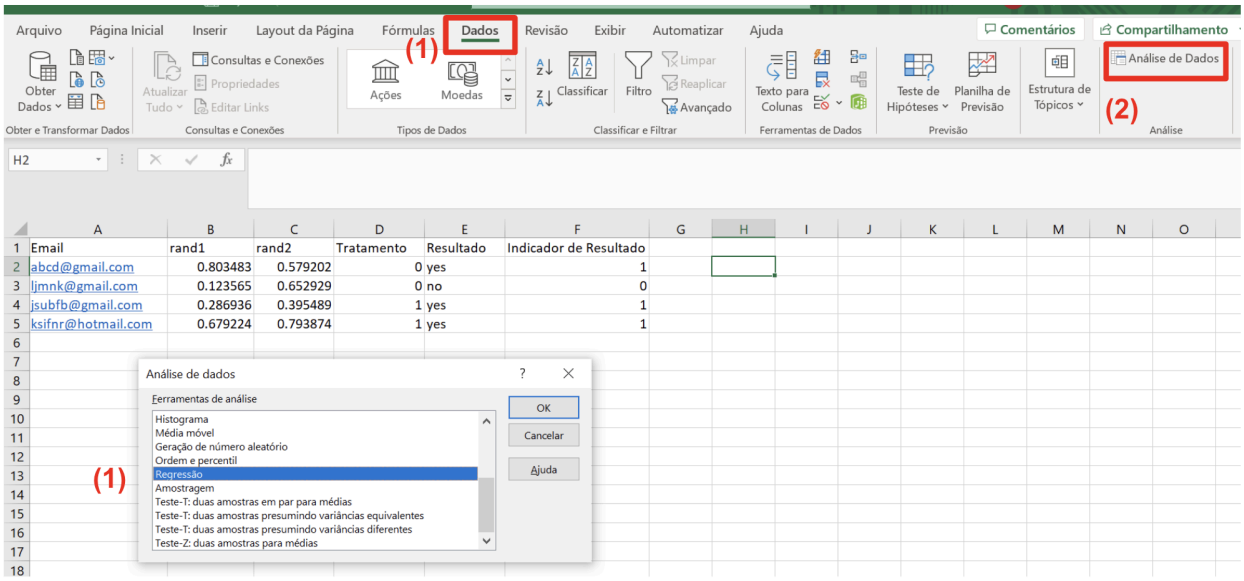
Resultado		Descrição
Média	0.75	Avanço de valores
Erro Padrão	0.25	Uma medida que lhe diz como a média da amostra se aproxima da média da população. Quanto maior o número, menor é a probabilidade de que a média da amostra seja precisa
Mediana	1	Ponto médio dos valores
Modo	1	Valor repetido com mais freqüência
Deviação Standard	0.5	A quantidade de variação de um conjunto de valores. Ou seja, quão semelhantes são os resultados entre unidades em sua amostra. Um exemplo em que a duração média do tempo que alguém leva para pagar uma multa é de 60 dias. Esta média pode ter um pequeno desvio padrão se -68% das pessoas pagarem sua multa entre 50 e 70 dias, ou um grande desvio padrão se -68% das pessoas pagarem sua multa entre 10 e 110 dias. Um desvio padrão maior indica uma maior "dispersão".
Variância de amostra	0.25	
Curtose	4	
Skewness	-2	
Gama	1	Diferença entre o valor mais alto e o valor mais baixo
Mínimo	0	O valor mais baixo
Máximo	1	Os valores mais altos
Soma	3	Soma de todos os valores
Contagem	4	Número de valores

Passo 2. Executar uma regressão

Na guia “Dados” (1), use a função “Análise de dados” (2) e selecione “Regressão” (3) como ferramenta de regressão[2]. Clique em "OK". (Ver imagem 13.)

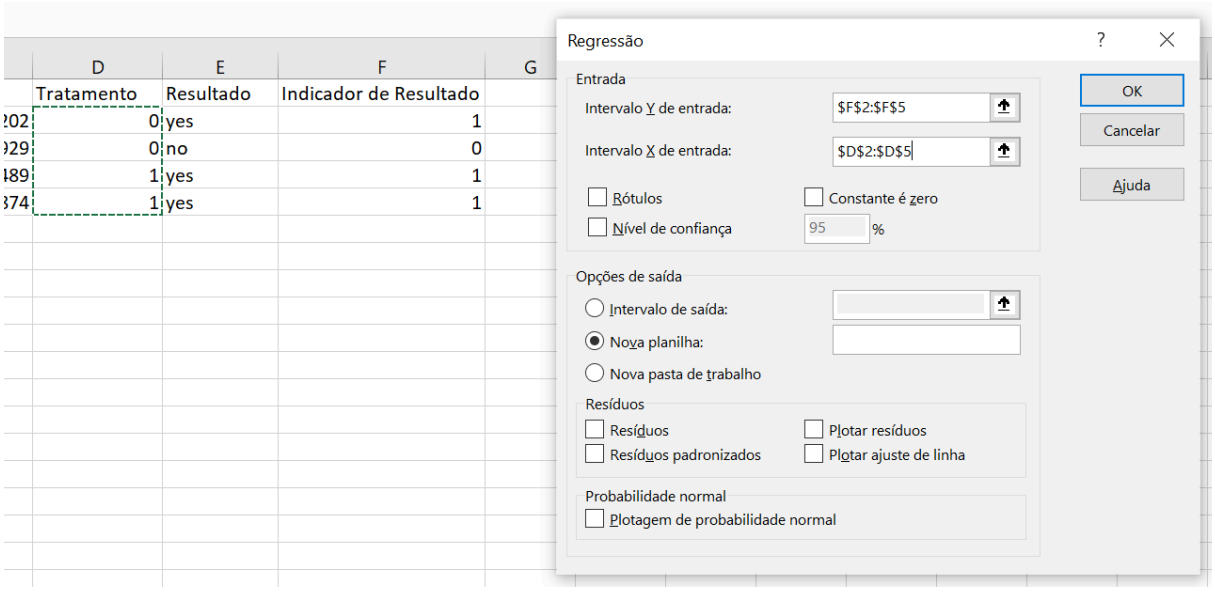
[2] *A regressão é um método estatístico que tenta determinar o caráter e a força da relação entre uma variável dependente – neste caso, cliques em um link – e uma ou múltiplas variáveis – neste caso, o recebimento de uma versão de uma carta. Em outras palavras, receber uma versão modificada da carta resulta em mais pessoas, menos pessoas ou a mesma quantidade de pessoas clicando em um link, em comparação com o recebimento da versão normal da carta?*

Imagem 13. Acessando a ferramenta de regressão



Na configuração da regressão, insira a coluna “indicador de resultado” como “Intervalo Y” e a coluna “Tratamento” como “Intervalo X”. Você pode a) escolher a coluna inteira e marcar “Rótulos” ou b) selecionar apenas os valores sob o rótulo da variável e certificar-se de que “Rótulos” esteja desmarcado. (Ver imagem 14.)

Imagem 14. Produzindo a regressão



Após pressionar em OK, o programa criará uma nova planilha com os dados de regressão semelhante à imagem 15.

Imagem 15. Dados de regressão

21	Resumo dos Resultados								
22	*Estatísticas de Regressão*								
23	Coeficiente de correlação múltipla	0.1389467							
24	R-quadrado	0.0164784							
25	R-quadrado ajustado	-0.0184383							
26	Erro Padrão	0.5968572							
27	Observações	30							
28									
29	ANOVA								
30		*df*	*SS*	*MS*	*F*	*Siginificancia F*			
31	Regressão	1	0.1333333	0.13333	0.5283	0.47339			
32	Residual	28	7.0666667	0.25238					
33	Total	29	7.2						
34									
35		Coeficientes	Erro Padrão	t-Stat	Valor-P	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
36	Intercepção	0.3333333	0.1297853	2.56893	0.01579	0.06783	0.569392	0.06739	0.59904
37	X Variável 1	0.13333333	0.18344853	0.72684	0.47895	-0.45256	0.57393	-0.24352	0.5091
38									

Passo 3. Interpretar os resultados da regressão

A tabela de regressão contém bastante informação, mas há 2 lugares importantes para olhar: 1) o número de observações (identificado na última linha da primeira tabela) e 2) a tabela inferior.

Você deve primeiro anotar o número de observações. É o que você esperava? Corresponde ao número de linhas em seus dados?

Em seguida, observe a tabela inferior, que contém seus principais resultados. A tabela deve ter 2 linhas: “Interceptação” e “X Variável”. A linha “Interceptação” contém informações sobre o **grupo do primeiro e-mail (geralmente o grupo de controle)** e a linha “X Variável” contém informações sobre o **grupo do segundo e-mail (geralmente o grupo de tratamento)**. Em nosso cenário hipotético, **estamos interessados em comparar a proporção de indivíduos no grupo do primeiro e-mail que clicam em um link com a proporção de indivíduos no grupo do segundo e-mail.**

A Tabela 1 inclui uma descrição de cada estatística relevante.

Tabela 1. Descrição dos dados estatísticos de regressão

Estatística	X Variável (E-mail 2)
Coeficiente	A diferença entre as médias dos grupos do primeiro e segundo e-mails, geralmente chamado de “efeito de tratamento”. Se observarmos que o coeficiente da “X Variável” é um número negativo, podemos interpretar que a média do segundo grupo é menor que a do primeiro grupo. Você pode multiplicar o número por 100 para ver esse valor como uma porcentagem. No exemplo acima, receber o segundo e-mail aumentaria a probabilidade de os participantes clicarem no link em 13,33%.
Erro padrão	Descreve a probabilidade de que a média do grupo do segundo e-mail da nossa

	amostra reflita com precisão a média da população. Ou seja, se 46,7% do grupo do segundo e-mail clicou no link, essa estatística nos diz a probabilidade de 46,7% de nossa população de interesse também clicar no link se recebesse esta versão do e-mail.
Valor-P	A probabilidade de identificar falsamente um efeito nos dados que não existe na realidade. A maioria dos experimentos chama os resultados de "estatisticamente significativos" quando há 5% ou menos de chance de que o efeito detectado não exista realmente.
Inferior 95%[3]	O limite inferior do intervalo de confiança de 95% do coeficiente para o grupo do segundo e-mail. Lembre-se de que o coeficiente representa a diferença entre os grupos do primeiro e segundo e-mail. Adicione a porcentagem desse valor ao coeficiente de interceptação * 100. Isso fornece o limite inferior do intervalo de confiança de 95% da média para o grupo do segundo e-mail.
Superior 95%[4]	O limite superior do intervalo de confiança de 95% do **coeficiente** para o grupo do segundo e-mail. Lembre-se de que o coeficiente representa a diferença entre os grupos do primeiro e segundo e-mail. Adicione a porcentagem desse valor ao coeficiente de interceptação * 100. Isso fornece o limite superior do intervalo de confiança de 95% da **média** para o grupo do segundo e-mail.

Calculando a média do grupo de controle

Se sua regressão incluir covariáveis, precisamos seguir alguns passos simples para calcular o resultado médio do grupo de controle.

Para calcular a média, suas variáveis "tratamento" e "indicador de resultado" devem ser formatadas como uma variável binária com zeros e uns. (Ver conceito 1 em limpeza de dados.)

Em sua planilha, insira o comando abaixo em qualquer célula vazia.

=MÉDIASE(D:D;"0";F:F)

[3] Um intervalo de confiança representa o intervalo no qual existe uma probabilidade especificada de que o valor de um determinado parâmetro (no nosso caso, a média de cliques) para a população de interesse esteja dentro desse intervalo. Uma probabilidade padrão para um intervalo de confiança é de 95%. Por exemplo, o intervalo de confiança de 95% para o grupo do primeiro e-mail é 6,7% - 59,9%. Isso significa que, se nossa população de interesse receber a primeira versão do e-mail, há 95% de probabilidade de que a média de cliques no link esteja dentro desse intervalo. O valor "95% inferior" e o valor "95% superior" representam os limites inferior e superior desse intervalo, respectivamente. Quanto menor a faixa do intervalo de confiança, mais confiantes podemos estar de que o verdadeiro valor de nossa variável de interesse para a população está próximo do valor observado da amostra.

[4] Ver nota acima.

H2 =MÉDIASE(D:D;"0";F:F)

	A	B	C	D	E	F	G	H	I
1	Email	rand1	rand2	treatment	outcome	outcome indicator			
2	abd123@gmail.com	0.6750148	0.5602122	0	yes	1		0.5	
3	klm021@gmail.com	0.2096577	0.6505005	0	no	0			
4	efg456@yahoo.com	0.088471	0.3785043	1	yes	1			
5	hij789@hotmail.com	0.1240896	0.7099048	1	yes	1			

Este comando faz com que o Excel execute um teste lógico: para todas as observações com valor “0” na coluna D (i.e., observações no grupo de controle), faça a média dos valores na coluna F.

Em outras palavras, o Excel faz três coisas. (1) O Excel conta o número de pessoas no grupo de controle. (2) O Excel obtém a soma dos resultados do grupo de controle. Como formatamos essa variável como zero ou um, a soma acaba sendo o número de pessoas no grupo de controle que clicaram no link. (3) O Excel obtém a soma dos resultados do grupo de controle e a divide pelo número de pessoas no grupo de controle.

Passo 4. Fazer um gráfico com os resultados da regressão

Podemos visualizar este resultado como um gráfico usando o “Graphing Template.xlsx” da BIT. Você precisará inserir várias informações dos dados da regressão (detalhadas acima) no modelo de gráfico para criar um gráfico. Você também precisará inserir a média do grupo de controle (detalhado acima). As células que você precisará referenciar dos dados de regressão e estatísticas descritivas estão destacadas na imagem 16 abaixo.

Imagem 16. Identificando os dados de regressão para o modelo gráfico

21	Resumo dos Resultados								
22	*Estatísticas de Regressão*								
23	Coeficiente de correlação múltipla	0.1389467							
24	R-quadrado	0.0164784							
25	R-quadrado ajustado	-0.0184383							
26	Erro Padrão	0.5968572							
27	Observações	30							
28									
29	ANOVA								
30		*df*	*SS*	*MS*	*F*	*Siginificancia F*			
31	Regressão	1	0.1333333	0.13333	0.5283	0.47339			
32	Residual	28	7.0666667	0.25238					
33	Total	29	7.2						
34									
35		Coeficientes	Erro Padrão	t-Stat	Valor-P	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
36	Intercepção	0.3333333	0.1297853	2.56893	0.01579	0.06783	0.569392	0.06739	0.59904
37	X Variável 1	0.13333333	0.18344853	0.72684	0.47895	-0.45256	0.57393	-0.24352	0.5091

Você precisará inserir esses valores nas células que correspondem a cada um deles no modelo de gráfico. A imagem 17 mostra um modelo preenchido e a Tabela 2 mostra onde encontrar cada valor nos dados da regressão e onde inseri-los no modelo de gráfico.

Imagem 17. Preenchendo o modelo de gráfico para representar graficamente os resultados da regressão

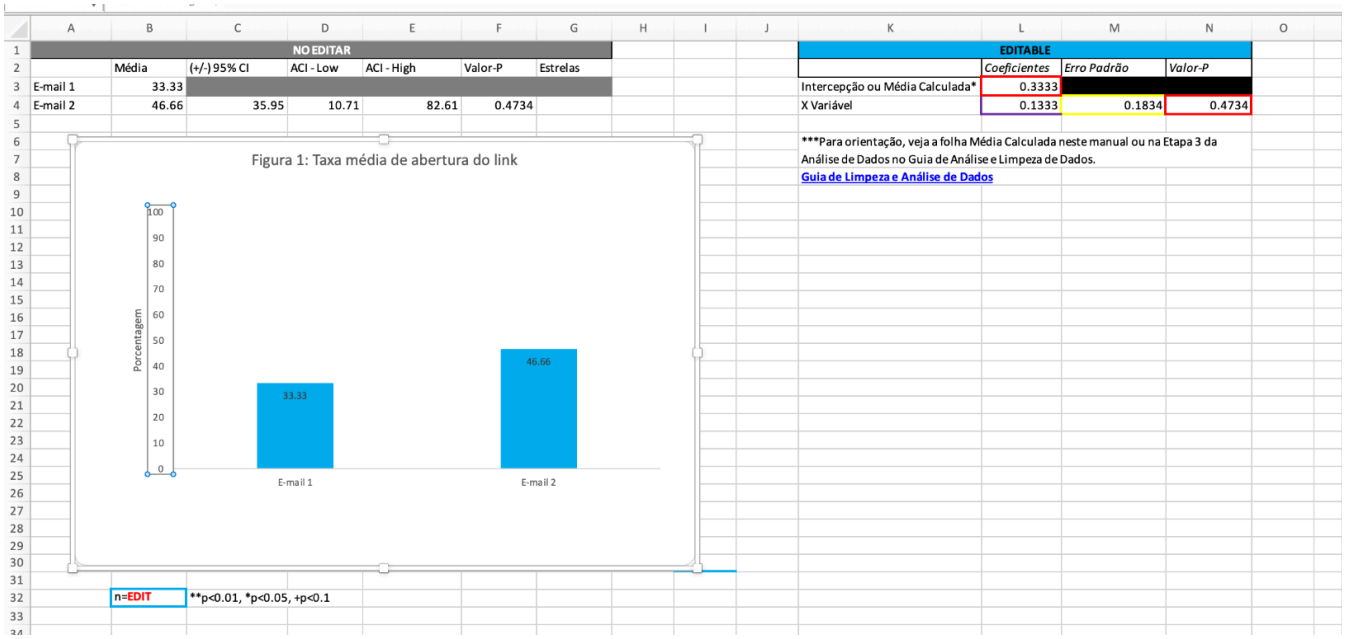


Tabela 2. Inserindo os valores-chave para produzir o gráfico

Valor	Onde encontrá-lo (Resultado da Regressão)					Onde escrevê-lo (Modelo gráfico)			
						EDITAR			
Média do primeiro e-mail (ou grupo de controle)		*Coeficientes*	*Erro Padrão*	*T-Statistic*	*P-valor*		*Coeficientes*	*Error Erro Padrã*	*P-valor*
	Intercepção	0.3333333	0.12971275	2.56978	0.01579	Intercepção	0.3333		
	Variável X 1	0.1333333	0.18344153	0.72684	0.47336	Variável X 1	0.1333333	0.18344153	0.47336
						EDITAR			
Efeito do tratamento		*Coeficientes*	*Erro Padrão*	*T-Statistic*	*P-valor*		*Coeficientes*	*Erro Padrão*	*P-valor*
	Intercepção	0.3333333	0.12971275	2.56978	0.01579	Intercepção	0.3333		
	Variável X 1	0.1333333	0.18344153	0.72684	0.47336	Variável X 1	0.1333333	0.18344153	0.47336
						EDITAR			
Erro Padrão		*Coeficientes*	*Erro Padrão*	*T-Statistic*	*P-valor*		*Coeficientes*	*Erro Padrão*	*P-valor*
	Intercepção	0.3333333	0.12971275	2.56978	0.01579	Intercepção	0.3333		
	Variável X 1	0.1333333	0.18344153	0.72684	0.47336	Variável X 1	0.1333333	0.18344153	0.47336
						EDITAR			
P-Valor		*Coeficientes*	*Erro Padrão*	*T-Statistic*	*P-valor*		*Coeficientes*	*Erro Padrão*	*P-valor*
	Intercepção	0.3333333	0.12971275	2.56978	0.01579	Intercepção	0.3333		
	Variável X 1	0.1333333	0.18344153	0.72684	0.47336	Variável X 1	0.1333333	0.18344153	0.47336

Toques finais

Depois de concluir o preenchimento dessas células, podemos trabalhar nos toques finais:

- **Altere o título do gráfico.** Clique duas vezes no título do gráfico e digite o nome apropriado.

- **Ajuste o alcance do eixo vertical.** Dependendo das médias de seus grupos e do intervalo de confiança, você pode querer ajustar o alcance do eixo vertical. Por exemplo, se a média de um grupo for 5,5% e do outro grupo for 10,5%, seria difícil visualizar a diferença se o eixo vertical se estendesse de 0 a 100%. Pode ser mais razoável ter um intervalo de 0 a 20%. (Ver imagem 18 para ver as etapas abaixo.)
    1. Clique no eixo vertical.
    2. A janela 'Formatar Eixo' deve aparecer.
    3. Em 'Opções de Eixo', altere o limite máximo. (*Para o exemplo acima, esse valor seria 20.*)
    4. Se desejar, altere a unidade principal. A unidade principal representa o espaço entre cada marca no eixo vertical. Neste modelo, a unidade principal é definida como 10. (*Para o exemplo acima, pode fazer sentido alterar para 4 ou 5.*)
- **Atualize o tamanho da amostra.** Clique duas vezes onde diz "n = EDIT" na nota de rodapé. Substitua "EDIT" pelo valor na célula "Observações" da tabela de saída destacada em azul. Certifique-se de alterar a cor da fonte do tamanho da amostra para preto. (Ver imagem 19 abaixo.)

Imagem 18. Ajuste da faixa do eixo vertical

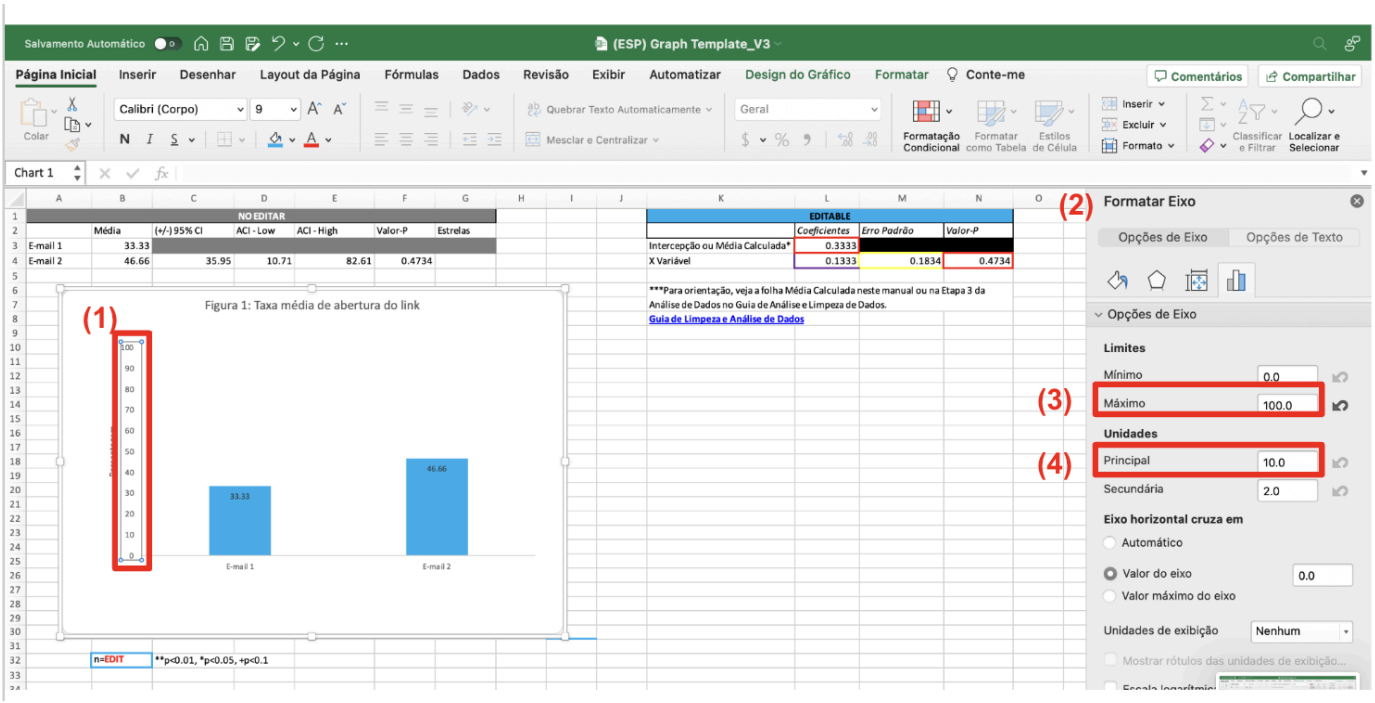
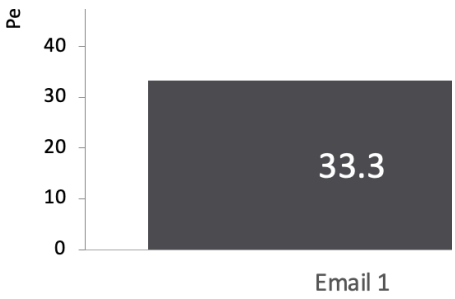


Imagem 19. Atualizando o tamanho da amostra



n = EDIT ** p<0.01, * p<0.05, + p<0.1

20		
21	Resumo dos Resultados	
22	*Estatísticas de Regressão*	
23	Coeficiente de correlação múltipla	0.1389467
24	R-quadrado	0.0164784
25	R-quadrado ajustado	-0.0184383
26	Erro Padrão	0.5968572
27	Observações	30

O gráfico agora está totalmente pronto para ilustrar o resultado do estudo no relatório.