Excel Guide: Cleaning & Analysis

There are two major steps for performing analysis: data cleaning and data analysis. This document will walk you through these steps using Microsoft Excel.[1] For this exercise, we are cleaning and analyzing the results of a hypothetical evaluation to see how a modified version of an email affects the number of recipients who click on a hyperlink in the email. In this hypothetical, treatment individuals received the modified email and control individuals received the business-as-usual email.

Pre-Cleaning

Before working with a dataset, you should create a copy of your raw data and review the data to understand what information the set contains and what changes you might need to make.

Creating a copy of raw data
When working with data, it is best practice to preserve all information in an unedited data file. Before performing any data cleaning or analysis, save the raw data (i.e. original, unedited data), then make a copy of the file for you to edit. This prevents data from being lost that we might need for subsequent analyses or to check that we have not made errors (e.g., unintentionally replacing an age value of 45 with 2).

Reviewing data
Spreadsheets often contain many columns of information and may include codes or shorthand (e.g., “Female” might be coded as “F”, 0, or 1), and fields may not be formatted correctly. Before cleaning your dataset, take a look at your data to understand what information the dataset contains and identify what you might need to clean.

When reviewing data, ask yourself the following questions:

- How many total observations (rows) are in my dataset?
- What does each observation (row) represent (e.g. a person, a household, a school)?
- Are there columns for all the data I need for analysis / that I expected to have?
- Which columns do I need for this analysis? Which columns do I not need that I can delete?
- Do I understand what variable each column represents?
    - *Ex: med_inc = Median income*
- Do I know what the values in each column mean?
    - *Ex: For a race column with coded values such as 1, 2, 3, 4, etc., which race does each value correspond to?*
- Are values formatted consistently within each column?

[1] This guide was created using Microsoft Excel 2021 Version 16.57 and the keyboard commands are defined for Windows keyboards. Locations of Excel functions may differ depending on the version of Excel that you are using, and keyboard commands will differ for other keyboards.

    - *Ex: Are names in all uppercase, lowercase, proper case, or a mix?*
- Are values formatted in a convenient way for analysis? If not, how should they be formatted? *(See Concept 1 under Data Cleaning for an example.)*
- Are there duplicate observations? *(See Concept 5 under Data Cleaning.)*
- Are there missing values for important columns? *(See Concept 4 under Data Cleaning.)*

Data Cleaning

After reviewing your data, you can delete columns that are unnecessary for your analysis. In the rest of this document we will review a few concepts for performing data cleaning in Excel.

What will you learn how to do?

1. Create numeric variables from text variables
2. Standardize capitalization
3. Remove / replace characters
4. Check for missing values
5. Check for duplicate observations
6. Remove duplicate observations

Concept 1. Create numeric variables from text variables

It's important to confirm that treatment status and outcome variables are indicated as zeros and ones. For example, if you are measuring whether email recipients clicked on a hyperlink, those who click the link should be indicated as "1" whereas those who do not should be indicated as "0" on Excel.

If your treatment status or outcome variable is not already recorded in zeros and ones (for example, it is common for the outcome to be recorded as "yes" or "no"), Excel can quickly convert them into zeros and ones with the following command:

=IF(E2="yes",1,0)

Step 1a. Next to the column labeled "outcome", label this column "outcome indicator", and in cell F2, enter the above command.

Figure 1. Assigning an outcome indicator value

	A	B	C	D	E	F
1	Email	rand1	rand2	treatment	outcome	outcome indicator
2	abd123@gmail.com	0.128533	0.900317651	0	yes	=IF(E2="yes",1,0)
3	klm012@gmail.com	0.2338778	0.690948625	0	no	
4	efg456@yahoo.com	0.5686872	0.020733377	1	yes	
5	hij789@hotmail.com	0.8373926	0.255109258	1	yes	

This command makes Excel perform a logical test: if the value of *E2* (Column E is the field that records the outcome in this example) is "yes", then indicate "1", and otherwise indicate "0".

Step 1b. Now, fill this command down the entire column to convert all outcomes to zeros and ones. You can do so manually or using keyboard commands.

Manual fill: Select the cell where you have entered the command (in this case, E2) and drag the fill handle  (the dot that appears in the bottom right corner) down to the final row.

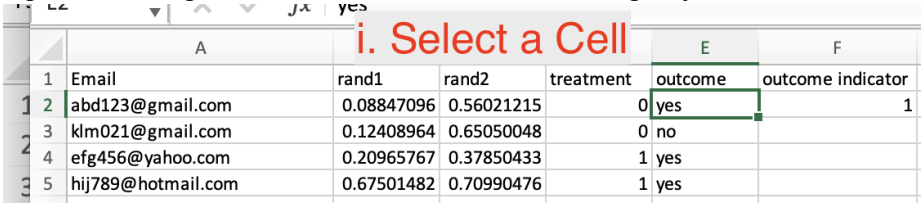
Figure 2. Manually filling a command down a column

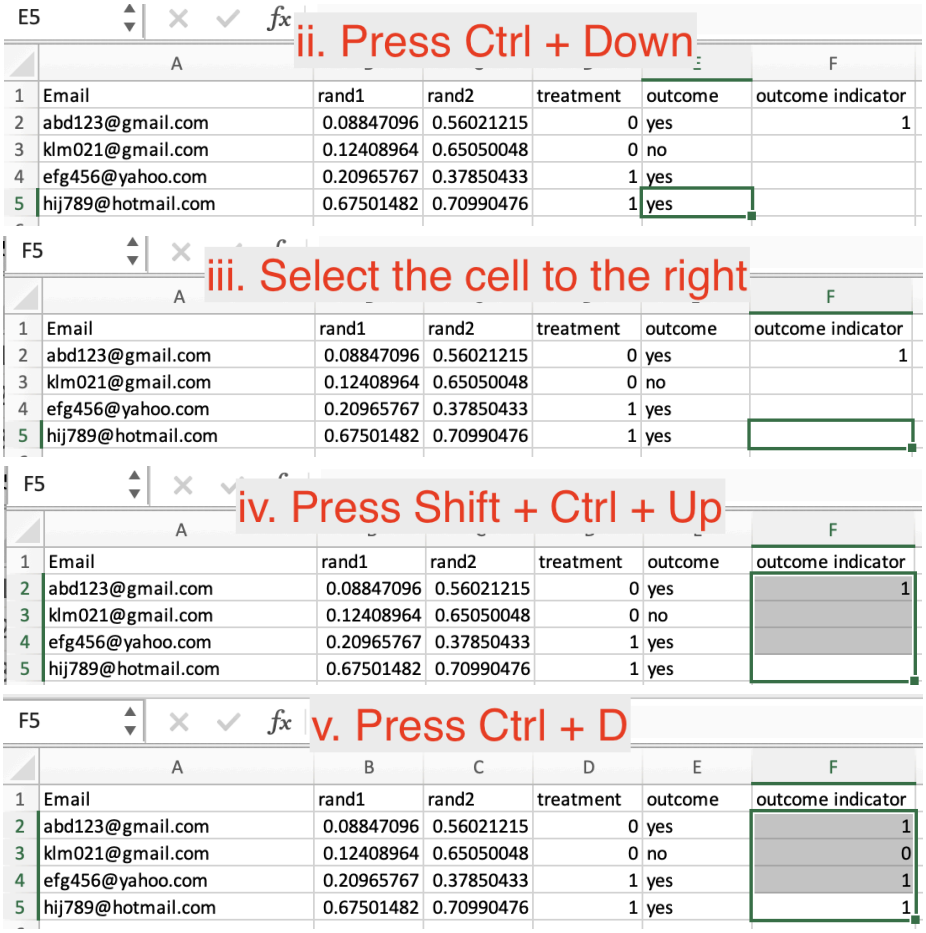
	A	B	C	D	E	F
1	Email	rand1	rand2	treatment	outcome	outcome indicator
2	abd123@gmail.com	0.128533	0.900317651	0	yes	1
3	klm012@gmail.com	0.2338778	0.690948625	0	no	0
4	efg456@yahoo.com	0.5686872	0.020733377	1	yes	1
5	hij789@hotmail.com	0.8373926	0.255109258	1	yes	1

Fill using keyboard commands. If the dataset is so large that it is difficult to manually scroll down to the end, you can also fill this command down the entire column using keyboard commands. (See Figure 3 below.)

i. Select a cell in a filled column, ideally the column next to the one you want to fill (in this case, "outcome").
ii. Press Ctrl + Down arrow. This brings you to the final row of your dataset.
iii. Select the cell to the right of the final cell of the "outcome" column. This will be the final cell of the "outcome indicator" column.
iv. Press Shift + Ctrl + Up arrow. This selects from the final cell up to the closest filled cell. In this case, E2 which contains our command.
v. Fill down the command by pressing Ctrl + D.

Figure 3. Filling a command down a column using keyboard commands





	Email	rand1	rand2	treatment	outcome	outcome indicator
2	abd123@gmail.com	0.08847096	0.56021215	0	yes	1
3	klm021@gmail.com	0.12408964	0.65050048	0	no	0
4	efg456@yahoo.com	0.20965767	0.37850433	1	yes	1
5	hij789@hotmail.com	0.67501482	0.70990476	1	yes	1

After filling down the command, we have a new column of zeros and ones that Excel can readily use for analysis.

Concept 2. Standardize capitalization

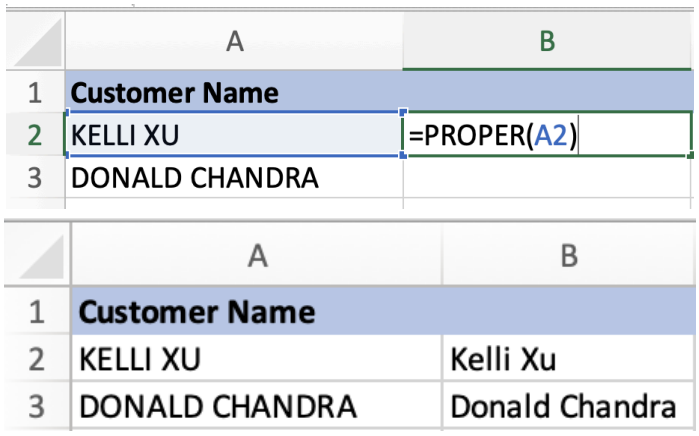
This is especially important for proper nouns. Later, we will learn how to check for and remove duplicate entries. It is important to standardize capitalization first, before removing duplicates. To understand why, imagine there are duplicate entries for a person named Kelli Xu but the capitalization differs (e.g., “Kelli Xu” vs. “KELLI XU”), Excel will not treat these as duplicate values until you standardize them.

Change the appearance of text using the following commands:

- UPPER() - converts text to all uppercase letters
- LOWER() - converts text to all lowercase letters
- PROPER() - converts text so the first letter of each word is uppercase; the rest, lowercase

Note: For names, we want to use “proper” capitalization (use the function “PROPER()”). Fill this command down the column by using the technique described above for assigning outcome indicator values.

Figure 4. Changing uppercase values to proper capitalization



	A	B
1	**Customer Name**	
2	KELLI XU	=PROPER(A2)
3	DONALD CHANDRA	

	A	B
1	**Customer Name**	
2	KELLI XU	Kelli Xu
3	DONALD CHANDRA	Donald Chandra

Concept 3. Remove / replace characters

Step 3a. Remove extra spaces & non-print characters. This step is also important to do before removing duplicates. Imagine if the duplicate value for Kelli Xu contains an erroneous space at the end of her name (e.g., “Kelli Xu” vs. “Kelli Xu ”), Excel will not treat these as duplicate values.

Step 3b. Replace characters with accents with the unaccented character. For example, Excel will not treat “Maria Ramirez” as a duplicate value of “María Ramírez”.

Identify, remove, and replace characters using the following commands / functions:

- Use the “TRIM()” function to remove extraneous spaces from text.
- Use “**Find & Replace**” to identify specific characters and replace them with other characters (e.g. find “ñ” and replace with “n”).

Concept 4. Check for missing values

There are often missing values in datasets. Sometimes, these values are missing for variables that are less consequential for our analysis. For example, it’s okay if we do not have middle names of some of the recipients of our emails, because knowing their middle names does not affect our understanding of the effect of the emails on recipients clicking on a link. On the other hand, if we are unsure which version of our email a group of recipients received because treatment assignment data is missing, this complicates our ability to identify the more effective version.

A good way to check for missing values is to use conditional formatting to highlight them.

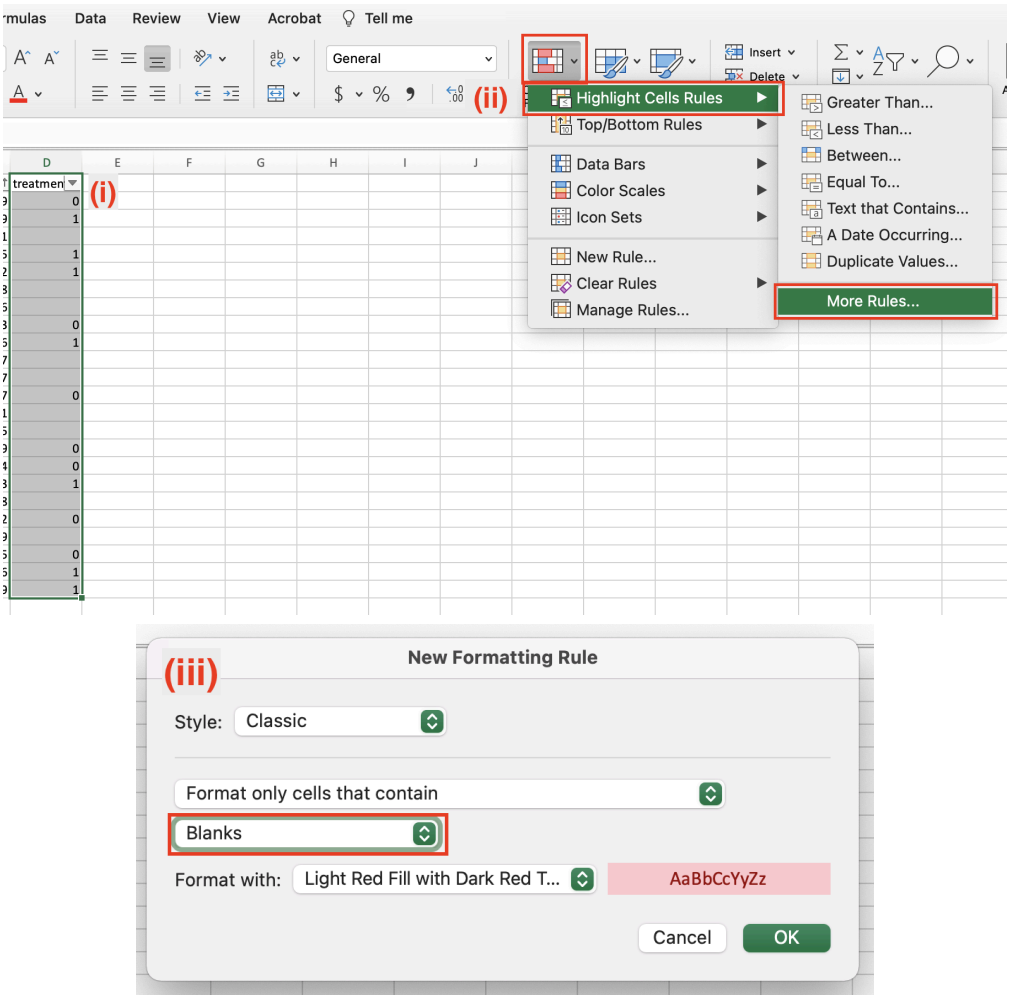
To highlight missing values (See Figure 5.)

i. Select the range of values that you would like to check. This could be for one variable, multiple variables, or your entire dataset. *Do not select the entire column(s), or else all empty cells below your dataset will be highlighted.*

ii. In the Home tab, select Conditional Formatting >> Highlight Cell Rules >> More Rules…

iii. In the box below “Format only cells that contain”, select “Blanks”, and click OK. All empty cells should be highlighted.

Figure 5. Conditional formatting for missing values



	A	B	C	D
1	Email	rand1	rand2	treatmen
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

To sort missing values (i.e. bring observations with missing values to the top of your dataset)

i. Select all values in your dataset.
ii. In the Home tab, select Sort & Filter >> Filter.
iii. Click the down arrow on the column label of interest.
iv. In the “By color” menu under Sort, choose Cell Color and the highlighted cell format.

To filter missing values (i.e. view only observations with missing values)
Note: when you filter out observations, they are hidden, not deleted.

i. Select all values in your dataset.
ii. In the Home tab, select Sort & Filter >> Filter.
iii. Click the down arrow on the column label of interest.
iv. In the “By color” menu under Filter, choose Cell Color and the highlighted cell format. You should only see observations with missing values in this column.

Note: To clear conditional formatting, in the Home tab, select Conditional Formatting >> Clear Rules >> Clear Rules from Entire Sheet or Clear Rules from Selected Cells.

Concept 5. Check for duplicate observations

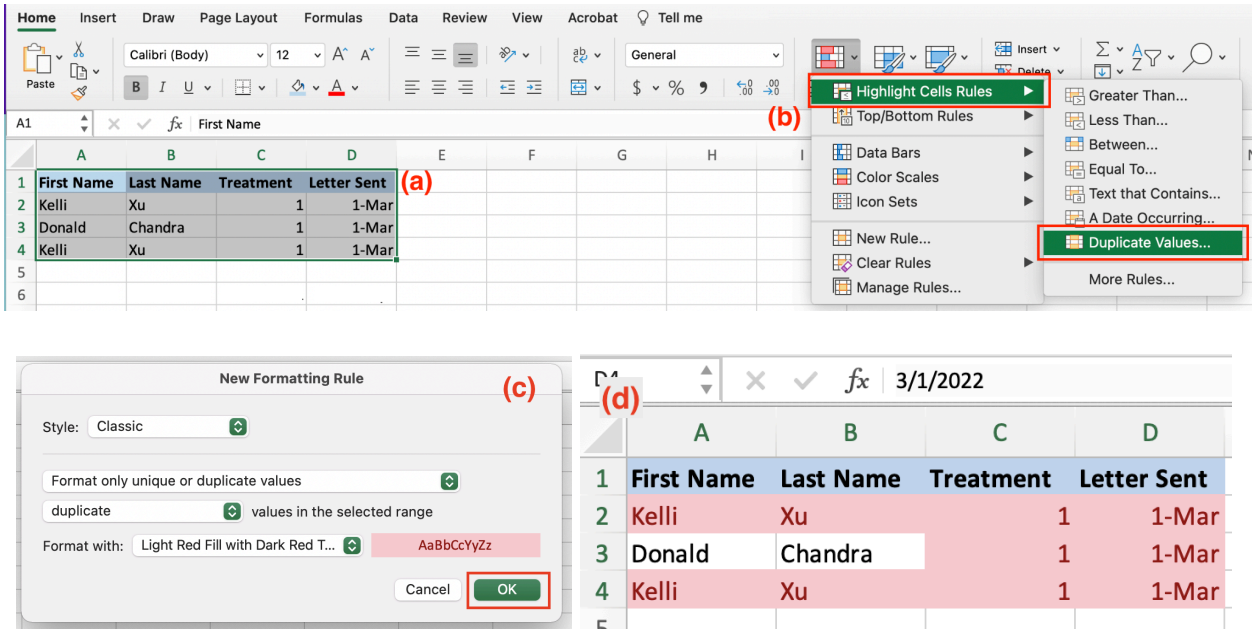
Check names and addresses to see if there are duplicates. If there are completely identical duplicate rows, delete them.

Let’s take for example a dataset where each observation is a person, and we have their full name. There will be lots of duplicate first names and there might be duplicate last names, so usually you’re trying to identify whether or not the *full name* is a duplicate. If the dataset already includes a column with a person’s full name, choose this variable when finding or removing duplicates. If one column includes a person’s first name and another column includes their last name, choose both of these columns when checking for duplicates. (For the steps below, see Figure 6.)

Step 5a. Select your data.

Step 5b. In the Home tab, select Conditional Formatting >> Highlight Cell Rules >> Duplicate Values…
Step 5c. In the New Formatting Rule window, click OK. (You can change the color of the highlighting in the "Format with" drop-down menu.
Step 5d. Duplicate values will now be highlighted.

Figure 6. Checking for duplicate values



(!) In Concept 6, you will learn how to remove duplicate observations, but let's think about whether duplicate observations make sense or if we should delete them.

In the example above, we see that there are identical observations for Kelli Xu. In a scenario in which our intervention involves sending one letter per person, we could conclude that we should delete the duplicate observation for Kelli Xu. We wouldn't want to count her outcomes twice when analyzing the data.

However, let's imagine an intervention in which one group of participants receive a letter and the other group receives the same letter AND a follow-up reminder letter. After highlighting duplicate values, our dataset could look something like Figure 7 below.

We would want to delete row 4 which is an exact duplicate of row 2, but we would not want to delete the 3rd observation for Kelli Xu in row 5, since this line shows when her second letter was sent.

Figure 7. Checking for duplicate values, continued

	A	B	C	D
1	First Name	Last Name	Treatment	Letter Sent
2	Kelli	Xu	1	1-Mar
3	Donald	Chandra	0	1-Mar
4	Kelli	Xu	1	1-Mar
5	Kelli	Xu	1	1-Apr

The key is to think about your intervention, then ask yourself: “Does it make sense that there are duplicate values?”

Concept 6. Remove duplicate observations

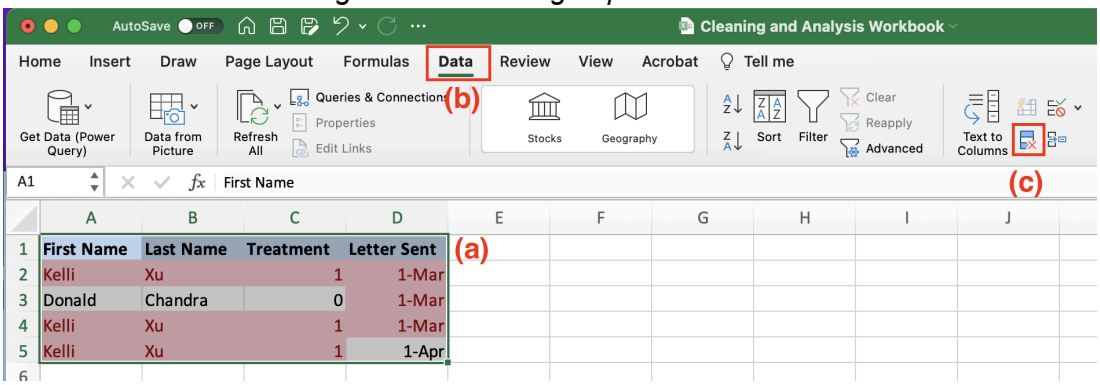
If you think that a duplicate observation should be deleted, follow the steps below to remove them. (See Figure 8.)

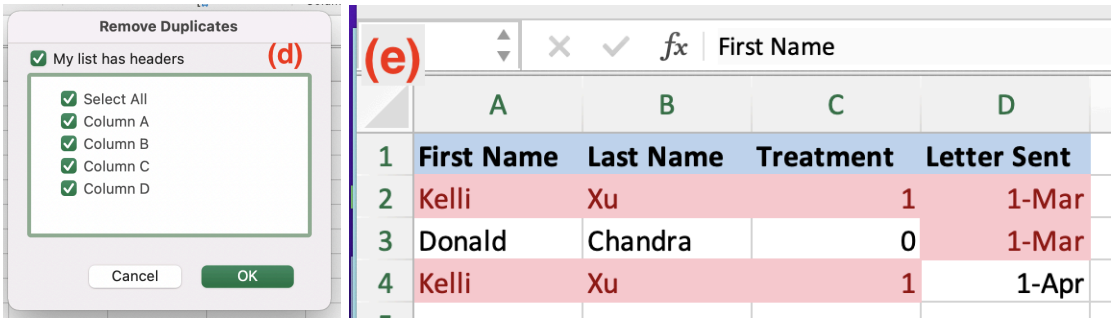
Step 6a. Select your data.
Step 6b. Go to the the Data tab
Step 6c. Click the Remove Duplicates button.
Step 6d. In the Remove Duplicates window, make sure “My list has headers” is checked if you selected all data including your headers. Then make sure that all columns are selected. Click OK.
Step 6e. Duplicate observations will now be deleted.

(!) Notice that only the duplicate entry for Kelli Xu receiving the letter on March 1st was removed, not the entry indicating that Kelli Xu received the letter on April 1st. That is because we selected all columns in the Remove Duplicates window. Selecting all columns means that the observation would have to be a duplicate across all of these columns.

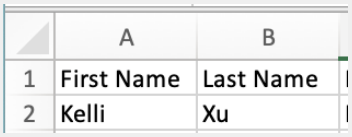
If we hadn’t selected Column D, Excel would have checked for observations that are identical for Columns A, B, and C. In that case, rows 4 and 5 would have been removed.

Figure 8. Removing duplicate entries





Some other functions that may be useful.
Strings are simply pieces of text that can be manipulated, queried, moved, and edited using additional standard Excel functions. For examples of the following functions, we will reference the data in the image below.



- CONCATENATE() – joins multiple strings together
  - Ex: =CONCATENATE(A2," ",B2) produces "Kelli Xu".
- LEFT() – returns the left n characters of a string
  - Ex: =LEFT(A2,3) produces "Kel".
- RIGHT() – returns the right n characters of a string
  - Ex: =RIGHT(A2,2) produces "li".
- MID() – returns characters from the middle of a string
  - Ex: =MID(A2,2,3) produces "ell".

Data Analysis

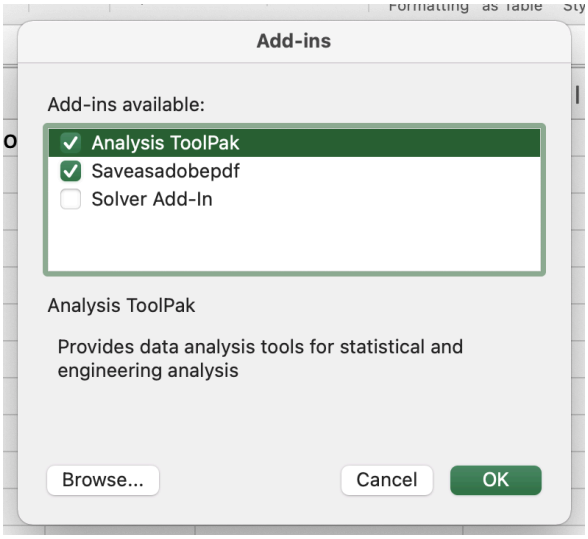
After cleaning your data, you are ready to start analyzing the data. Data analysis helps us to better understand our data, discover patterns, draw conclusions, and inform decision-making. This section provides a basic overview of how to describe data and draw conclusions from a basic regression analysis of a randomized controlled trial.

What will you learn how to do?

1. Produce and interpret descriptive statistics
2. Run a regression
3. Interpret regression outputs
4. Graph the regression results

To carry out the statistical analysis, we will use Excel's Analysis Toolpak Add-in (Installation Directions).

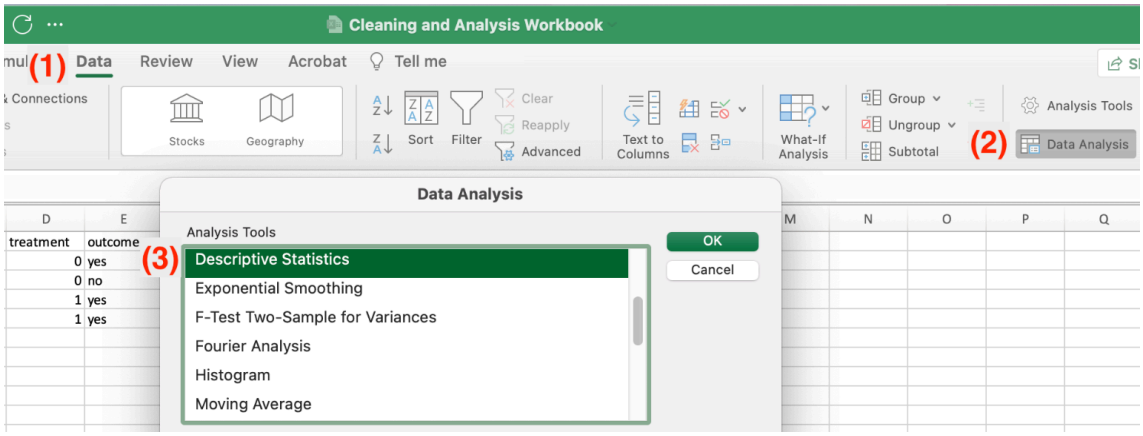
Figure 9. Microsoft Excel's Analysis Toolpak Add-in



Step 1. Produce and interpret descriptive statistics

After installing, under the "Data" tab (1), use the "Data Analysis" function (2), and select "Descriptive Statistics" (3). Click "OK". (See Figure 10.)

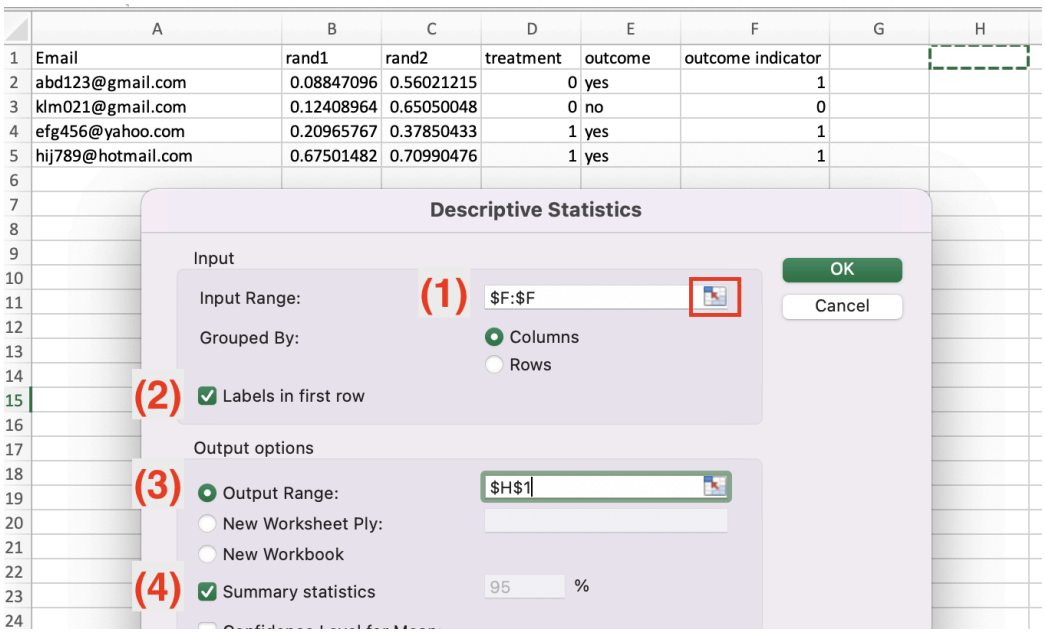
Figure 10. Accessing the Descriptive Statistics tool



In the “Descriptive Statistics” Window, select the data for the variable for which you want to produce descriptive statistics as the “Input Range” by clicking the button (1). You can either a) choose the entire column (in this case, by clicking on column F) and check “Labels in first row” (2) or b) select only the values under the variable label and make sure “Labels in first row” is unchecked.

Next, choose where you want the summary statistic table to appear, by clicking the button beside “Output Range” (3). You can choose wherever you would like for the table to appear, and the cell that you select will be the top left corner of the table. In this example, we have selected H1 as the beginning of the table. Finally, check the “Summary Statistics” option (4), and click “OK”. (See Figure 11.)

Figure 11. Producing the descriptive statistics



After you click “OK”, the summary statistics table for this variable will appear where you set the output range and will look similar to the first two columns of the table in Figure 12. In the third column, we have included descriptions for common concepts.

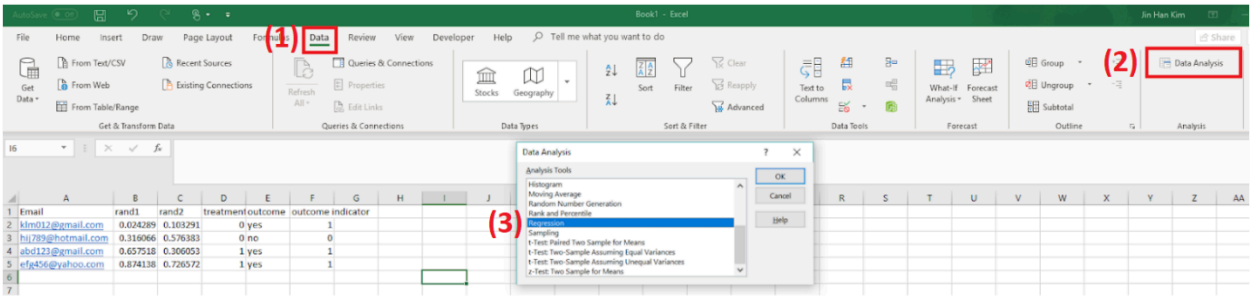
Figure 12. Summary statistics table with descriptions

outcome		*Description*
Mean	0.75	Average of values
Standard Error	0.25	A measure that tells you how precisely the sample mean approximates the population mean. The greater the number, the less likely it is that the sample mean is accurate.
Median	1	Midpoint of the values
Mode	1	Value repeated most frequently
Standard Deviation	0.5	The amount of variation of a set of values. That is, how similar outcomes are between units in your sample Take an example in which the average length of time it takes someone to pay a fine is 60 days. This average could have a small standard deviation if ~68% of people pay their fine between 50 and 70 days, or a large standard deviation if ~68% of people pay their fine between 10 and 110 days. A larger standard deviation indicates greater "dispersion".
Sample Variance	0.25	
Kurtosis	4	
Skewness	-2	
Range	1	Difference between the highest value and the lowest value
Minimum	0	Lowest value
Maximum	1	Highest value
Sum	3	Sum of all values
Count	4	Number of values

Step 2. Run a regression

Under the "Data" tab (1), use the "Data Analysis" function (2), and select "Regression" (3) as the regression tool.[2] Click "OK". (See Figure 13.)

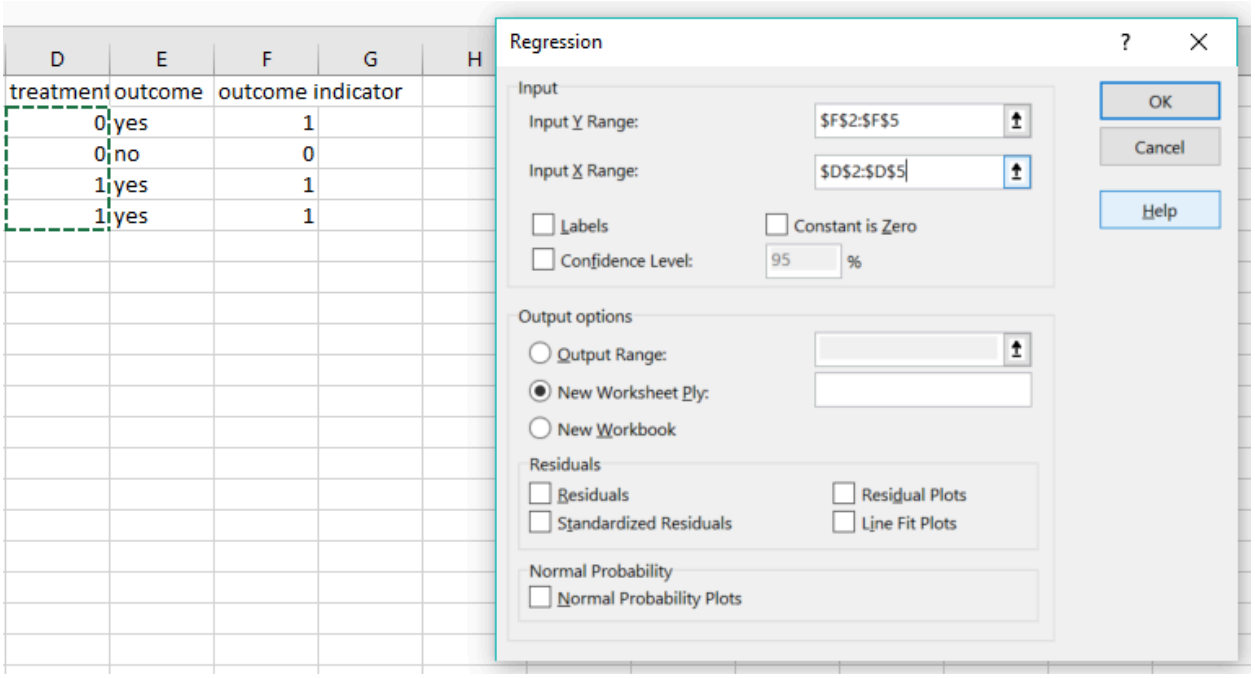
Figure 13. Accessing the Regression tool



[2] Regression is a statistical method that attempts to determine the character and strength of the relationship between one dependent variable–in this case, clicking on a link–and one or a series of other variables–in this case, receipt of a version of a letter. In other words, does receiving a modified version of a letter result in more people, less people, or the same amount of people to click on a hyperlink, compared to receiving the business-as-usual version of the letter?

In the regression set-up, input the “outcome indicator” column as the “Y Range” and the treatment column as the “X Range”. You can either a) choose the entire column and check “Labels” or b) select only the values under the variable label and make sure “Labels” is unchecked. (See Figure 14.)

Figure 14. Producing the regression



Once we press OK, the program will create a new worksheet with the regression output similar to Figure 15.

Figure 15. Regression Output

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.1360828							
5	R Square	0.0185185							
6	Adjusted R Squ	-0.0165344							
7	Standard Error	0.5023753							
8	Observations	30							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	0.13333333	0.13333	0.5283	0.47336			
13	Residual	28	7.06666667	0.25238					
14	Total	29	7.2						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	0.3333333	0.12971275	2.56978	0.01579	0.06763	0.59904	0.06763	0.59904
18	X Variable 1	0.1333333	0.18344153	0.72684	0.47336	-0.24243	0.5091	-0.24243	0.5091

Step 3. Interpret regression outputs

The regression table contains quite a bit of information, but there are 2 important places to look: 1) at the number of observations (identified in the bottom row of the first table) and 2) at the bottom table.

You should first note the number of observations. Is it what you expected? Does it match the number of rows in your data?

Next, look at the bottom table, which contains your key results. The table should have 2 rows: "Intercept" and "X Variable." The "Intercept" row contains information about the **first email group (often your control group)** and the "X Variable" row contains information about the **second email group (often your treatment group).** In our hypothetical scenario, **we are interested in comparing the proportion of individuals in the first email group who click a link to the proportion of individuals in the second email group.**

Table 1 includes a description of each relevant statistic.

Table 1. Description of regression output statistics

Statistic	X Variable (Email 2)
Coefficient	The difference between the first and second email group's averages, often called the "treatment effect". If we see that the coefficient of "X Variable" is a negative number, we can interpret that the second group's average is lower than that of the first group. You can multiply the number by 100 to see this value as a percentage. In the example above, receiving the second email would increase the likelihood of participants clicking on the link by 13.33%.

Standard Error	Describes how likely it is that the average of the second email group in our sample accurately reflects the average of the population. That is, if 46.7% of the second email group clicked the link, this statistic tells us how likely it is that 46.7% of our population of interest would also click the link if they received this version of the email.
P-value	The probability of falsely identifying an effect in your data that does not exist in reality. Most experiments call results "statistically significant" when there is a 5% or less chance that the effect they detected doesn't actually exist.
Lower 95%[3]	The lower bound of the 95% confidence interval of the coefficient for the second email group. Recall that the coefficient represents the difference between the first and the second email group. Add the percentage of this value to the intercept coefficient * 100. This gives you the lower bound of the 95% confidence interval of the average for the second email group.
Upper 95%[4]	The upper bound of the 95% confidence interval of the **coefficient** for the second email group. Recall that the coefficient represents the difference between the first and the second email group. Add the percentage of this value to the intercept coefficient * 100. This gives you the upper bound of the 95% confidence interval of the **average** for the second email group.

Calculating the control group mean

If your regression does include covariates, we need to take a few simple steps to calculate the mean outcome of the control group.

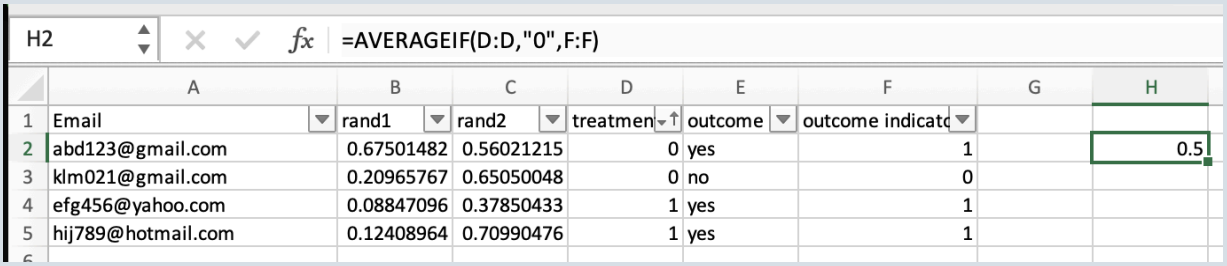
To calculate the mean, your "treatment" and "outcome indicator" variables must be formatted as a binary variable with zeros and ones. (See Concept 1 under Data Cleaning.)

In your spreadsheet, enter the command below into any empty cell.

=AVERAGEIF(D:D,"0",F:F)

[3] A confidence interval represents the range in which there is a specified probability that the value of a given parameter (in our case, the average clicks) for the population of interest lies within this range. A standard probability for a confidence interval is 95%. For example, the 95% confidence interval for the first email group is 6.7% - 59.9%. This means that if our population of interest were to receive the first version of the email, there is a 95% probability that the average of link clicks would lie within this range. The "Lower 95%" value and the "Upper 95%" value represent the lower and upper bounds of this range, respectively. The smaller the range of the confidence interval, the more confident we can be that the true value of our variable of interest for the population is close to the observed value of the sample.

[4] See above note.



	A	B	C	D	E	F	G	H
1	Email	rand1	rand2	treatmen	outcome	outcome indicato		
2	abd123@gmail.com	0.67501482	0.56021215	0	yes	1		0.5
3	klm021@gmail.com	0.20965767	0.65050048	0	no	0		
4	efg456@yahoo.com	0.08847096	0.37850433	1	yes	1		
5	hij789@hotmail.com	0.12408964	0.70990476	1	yes	1		

This command makes Excel perform a logical test: for all observations with a value of "0" in column D (i.e. observations in the control group), take the average of the values in column F.

In other words, Excel does three things. (1) Excel counts the number of people in the control group. (2) Excel takes the sum of outcomes for the control group. Because we have formatted this variable as either a zero or a one, the sum ends up being the number of people in the control group who clicked the link. (3) Excel takes the sum of the outcomes for the control group and divides it by the number of people in the control group.

Step 4. Graph the regression results

We can visualize this result as a graph by using BIT's "Graphing Template.xlsx". You will need to feed several pieces of information from the regression output (detailed above) into the graphing template to create a graph. You will also need to input the control group mean (detailed above). The cells you will need to reference from the regression output and descriptive statistics are highlighted in Figure 16 below.

Figure 16. Identifying regression output values for the graphing template

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.1360828							
5	R Square	0.0185185							
6	Adjusted R Squ	-0.0165344							
7	Standard Error	0.5023753							
8	Observations	30							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	0.13333333	0.13333	0.5283	0.47336			
13	Residual	28	7.06666667	0.25238					
14	Total	29	7.2						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	0.3333333	0.12971275	2.56978	0.01579	0.06763	0.59904	0.06763	0.59904
18	X Variable 1	0.1333333	0.18344153	0.72684	0.47336	-0.24243	0.5091	-0.24243	0.5091

You will need to enter these values into the cells that correspond to each of them in the graphing template. Figure 17 shows a completed template and Table 2 shows where to find each value in the regression output and where to input them in the graphing template.

Figure 17. Filling in the graphing template to graph the regression results

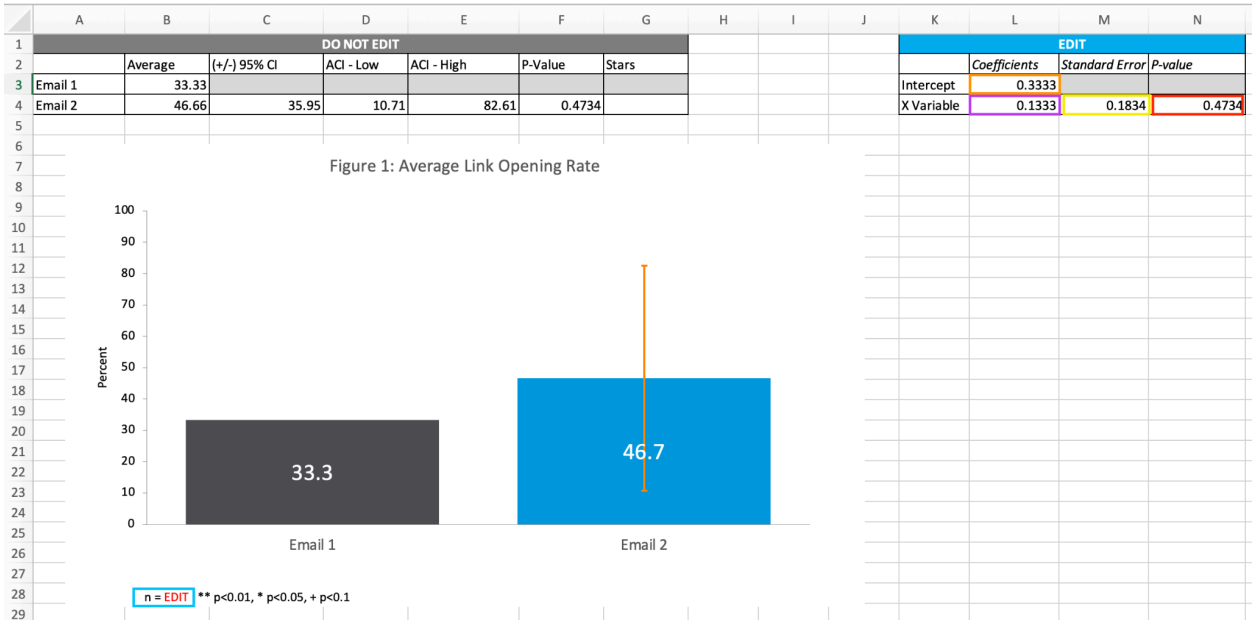
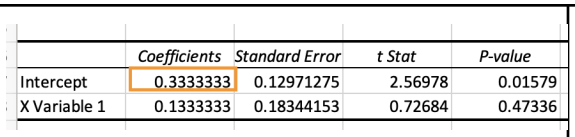
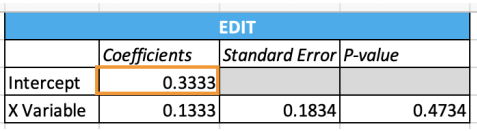
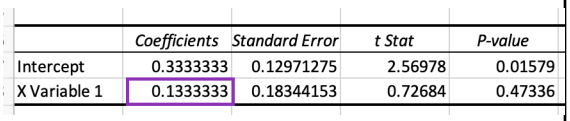
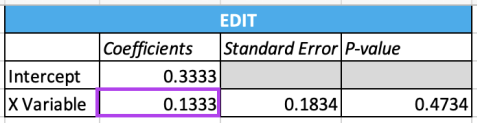
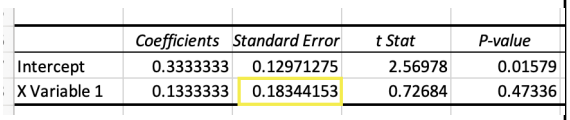
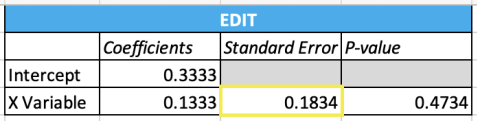
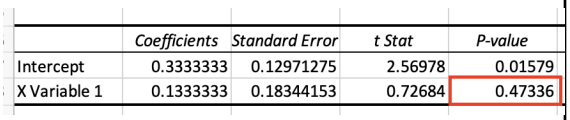
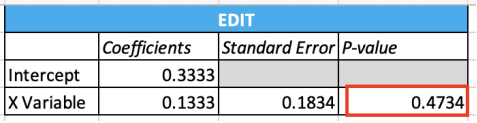


Table 2. Inputting key values to produce the graph

Value	Where you can find it (Regression Output)	Where to type it (Graphing Template)

Mean for the first email (or control group)		
Treatment effect		
Standard error		
P-value		

Finishing touches

Once we complete filling in these cells, we can then work on the final touches:

- **Change the graph title.** Double click the title of the graph and type in the appropriate name.
- **Adjust the range of the vertical axis.** Depending on the averages of your groups and the confidence interval, you may want to adjust the range of the vertical axis. For example, if the average of one group is 5.5% and the other group is 10.5%, it would be hard to visualize the difference if the vertical axis extends from 0 - 100%. It may be more reasonable to have a range from 0 - 20%. (See Figure 18 for steps below.)
    1. Click the vertical axis.
    2. The Format Axis window should appear.
    3. Under Axis options, change the maximum bound. *(For the example above, this value would be 20.)*
    4. If desired, change the major unit. The major unit represents the space between each tick mark on the vertical axis. In this template, the major unit is set as 10. *(For the example above, it might make sense to change this to 4 or 5.)*
- **Update the sample size.** Double click where it says “n = EDIT” in the footnote. Replace “EDIT” with the value on the output table’s “Observations” cell outlined in blue. Make sure to change the font color of the sample size to black. (See Figure 19 below.)

Figure 18. Adjusting the range of the vertical axis

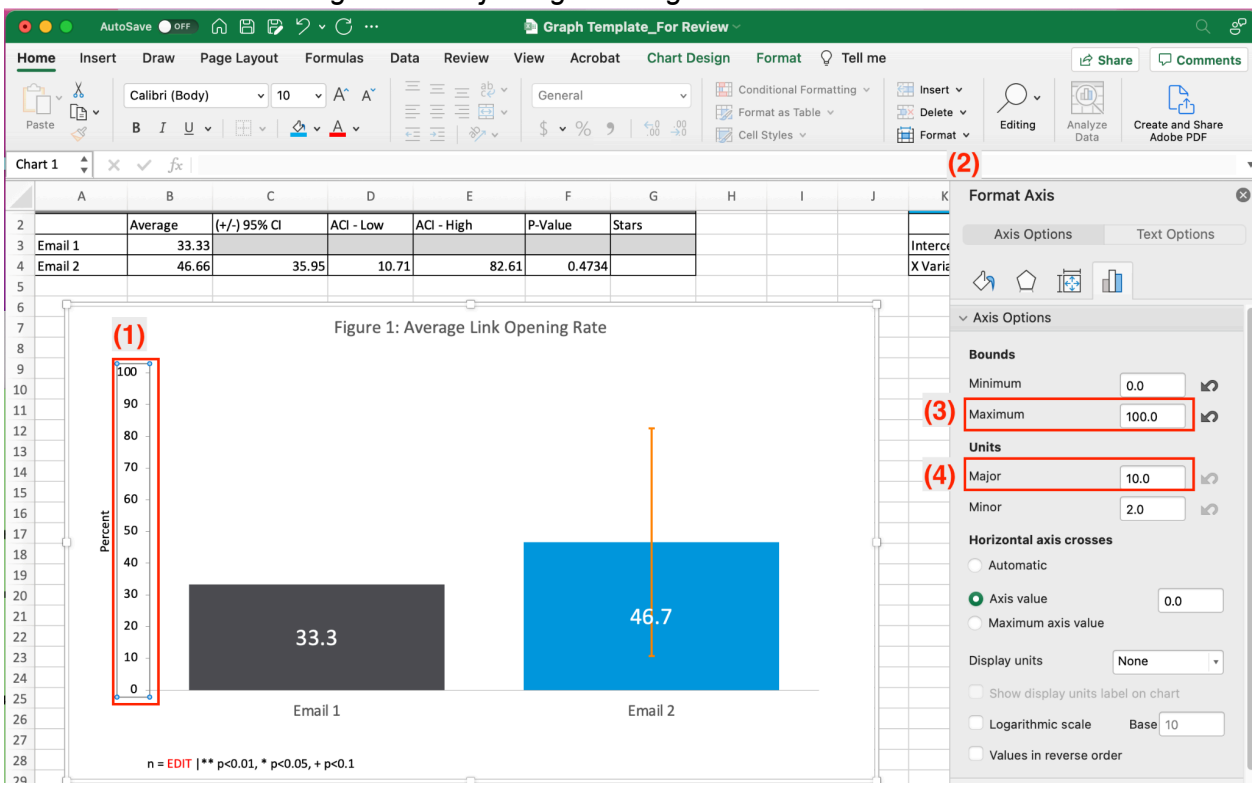
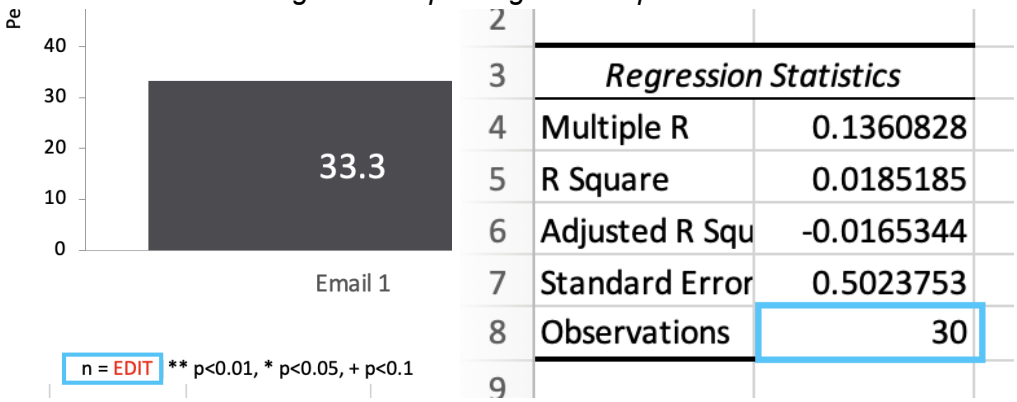


Figure 19. Updating the sample size



The graph is now fully ready to illustrate the trial result in the report.