Guía de Excel: Limpieza & Análisis

Hay dos pasos principales para realizar el análisis: la limpieza de datos y el análisis de datos. Este documento le guiará a través de estos pasos utilizando Microsoft Excel. Para este ejercicio, estamos limpiando y analizando los resultados de una evaluación hipotética para ver cómo una versión modificada de un correo electrónico afecta al número de destinatarios que hacen clic en un hipervínculo del correo electrónico. En este caso hipotético, los individuos del grupo de tratamiento recibieron el correo electrónico modificado y los miembros del grupo de control recibieron el correo electrónico como de costumbre.

Antes de la limpieza

Antes de trabajar con un conjunto de datos, usted debe crear una copia de los datos sin procesar y revisarlos para comprender qué información contiene el conjunto y qué cambios podría tener que hacer.

Creación de una copia de los datos sin procesar
Cuando se trabaja con datos, la mejor práctica es conservar toda la información en un archivo de datos sin editar. Antes de realizar cualquier tipo de limpieza o análisis, guarde los datos sin procesar (es decir, los datos originales, sin editar), y luego haga una copia del archivo para poder editarlo. Esto evita que se pierdan datos que podríamos necesitar para análisis posteriores o para comprobar que no hemos cometido errores (por ejemplo, sustituir involuntariamente un valor de edad de 45 por 2).

Revisión de los datos
Las hojas de cálculo suelen contener muchas columnas de información y pueden incluir códigos o abreviaturas (por ejemplo, "Mujer" puede estar codificada como "F", 0 o 1), y los campos pueden no estar formateados correctamente. Antes de limpiar su conjunto de datos, revíselos para entender qué información contiene el conjunto de datos e identificar lo que podría necesitar limpiar.

Cuando revise los datos, hágase las siguientes preguntas:

- ¿Cuántas observaciones totales (filas) hay en mi conjunto de datos?
- ¿Qué representa cada observación (fila) (p.ej., una persona, un hogar, una escuela)?
- ¿Hay columnas para todos los datos que necesito para el análisis / que esperaba tener?
- ¿Qué columnas necesito para este análisis? ¿Qué columnas no necesito que pueda eliminar?
- ¿Entiendo qué variable representa cada columna?
    - *Ej: med_ing = Mediana de Ingresos*
- ¿Sé qué significan los valores de cada columna?
    - *Ej: Para una columna de raza con valores codificados como 1, 2, 3, 4, etc., ¿a qué raza corresponde cada valor?*

- ¿Los valores están formateados de forma coherente dentro de cada columna?
    - *Ej.: ¿Los nombres están en mayúsculas, minúsculas, mayúsculas de nombres propios o una mezcla?*
- ¿Los valores están formateados de manera conveniente para el análisis? Si no es así, ¿cómo deberían formatearse? *(Véase el concepto 1 en Limpieza de datos para ver un ejemplo.)*
- ¿Hay observaciones duplicadas? *(Véase el Concepto 5 en Limpieza de datos)*
- ¿Faltan valores en columnas importantes? *(Véase el Concepto 4 en Limpieza de datos)*

Limpieza de Datos

Después de revisar sus datos, puede eliminar las columnas que sean innecesarias para su análisis. En el resto de este documento revisaremos algunos conceptos para realizar la limpieza de datos en Excel.

¿Qué aprenderá a hacer?

1. Crear variables numéricas a partir de variables de texto
2. Estandarizar las mayúsculas y minúsculas
3. Eliminar / sustituir caracteres
4. Comprobar si faltan valores
5. Comprobar si hay observaciones duplicadas
6. Eliminar las observaciones duplicadas

Concepto 1. Crear variables numéricas a partir de variables de texto

Es importante confirmar que el estado del tratamiento y las variables de resultado se indican como ceros y unos. Por ejemplo, si está midiendo si los destinatarios del correo electrónico han hecho clic en un hipervínculo, los que hacen clic en el enlace deben indicarse como "1", mientras que los que no lo hacen deben indicarse como "0" en Excel.

Si su estado de tratamiento o variable de resultado no se registra ya en ceros y unos (por ejemplo, es habitual que el resultado se registre como "sí" o "no"), Excel puede convertirlos rápidamente en ceros y unos con el siguiente comando

=SI(E2=”si”,1,0)

Paso 1a. Junto a la columna denominada "resultado", etiquete esta columna como "indicador de resultado", y en la celda F2, introduzca el comando anterior.

Figura 1. Asignación de un valor de indicador de resultado

	A	B	C	D	E	F
1	**Email**	**rand1**	**rand2**	**tratamiento**	**resultado**	**indicador de resultado**
2	klm021@gmail.com	0.1067483	0.8221911	0	yes	=IF(E2="yes",1,0)
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						

Este comando hace que Excel realice una prueba lógica: si el valor de E2 (la columna E es el campo que registra el resultado en este ejemplo) es "sí", entonces indica "1", y en caso contrario indica "0".

Paso 1b. Ahora, complete este comando por toda la columna para convertir todos los resultados en ceros y unos. Puede hacerlo manualmente o mediante comandos de teclado.

Llenado manual: Selecciona la celda donde ha introducido la orden (en este caso, E2) y arrastre el tirador de relleno  (el punto que aparece en la esquina inferior derecha) hasta la última fila.

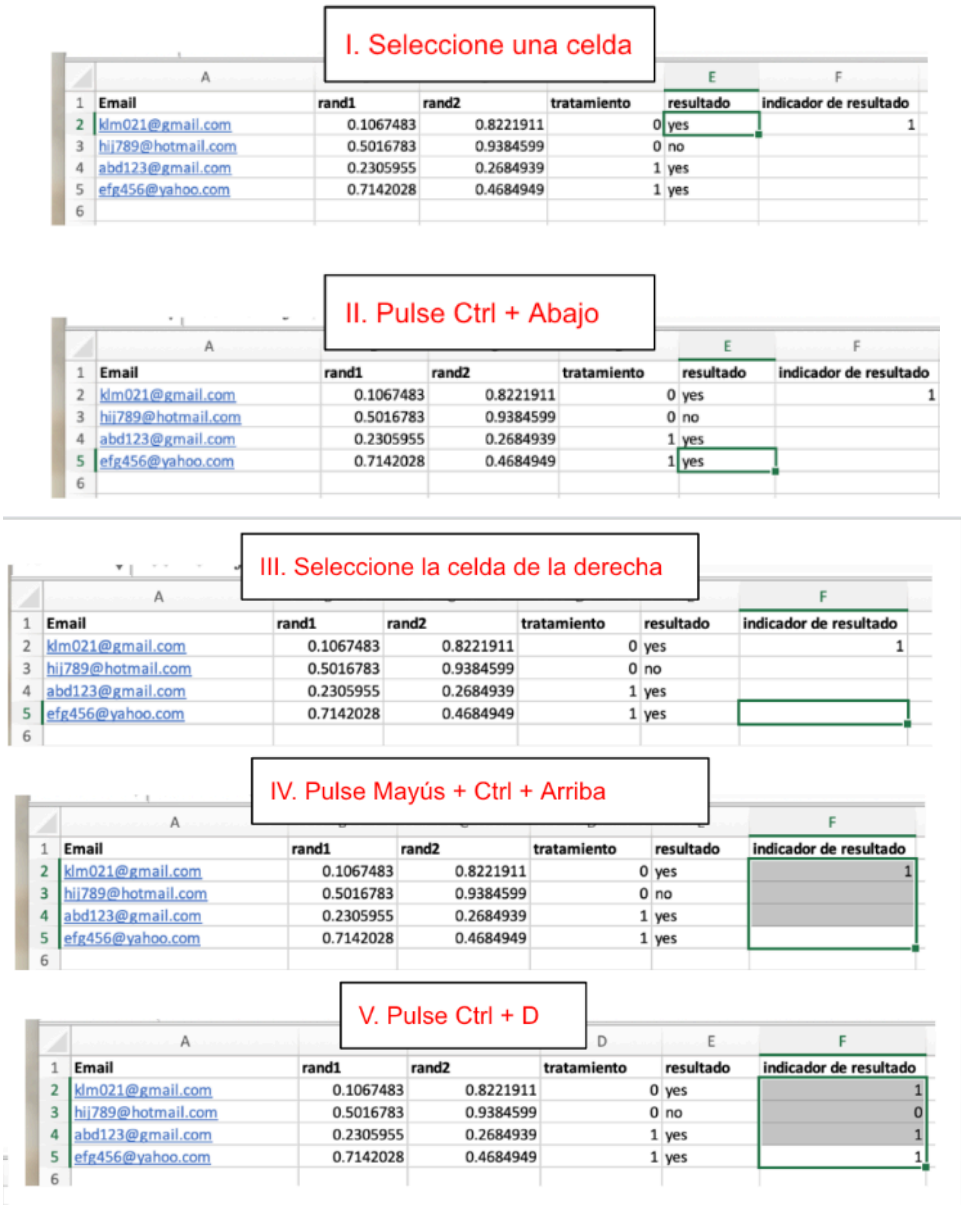
Figura 2. Rellenar manualmente un comando en una columna

	A	B	C	D	E	F
1	**Email**	**rand1**	**rand2**	**tratamiento**	**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						

Rellenar utilizando comandos de teclado. Si el conjunto de datos es tan grande que resulta difícil desplazarse manualmente hasta el final, también puede rellenar toda la columna mediante comandos de teclado. (Véase la Figura 3 más abajo).

i. Seleccione una celda en una columna rellenada, idealmente la columna siguiente a la que desea rellenar (en este caso, "resultado").
ii. Pulse Ctrl + Flecha abajo. Esto le llevará a la última fila de su conjunto de datos.
iii. Selecciona la celda a la derecha de la última celda de la columna "resultado". Esta será la última celda de la columna "indicador de resultados".
iv. Pulse Shift + Ctrl + Flecha arriba. Esto selecciona desde la última celda hasta la celda rellena más cercana. En este caso, E2 que contiene nuestro comando.
v. Rellene el comando pulsando Ctrl + D.

Figura 3. Rellenar una columna con comandos de teclado



	Email	rand1	rand2	tratamiento	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	

Después de rellenar el comando, tenemos una nueva columna de ceros y unos que Excel puede utilizar fácilmente para el análisis.

Concepto 2. Estandarizar las mayúsculas

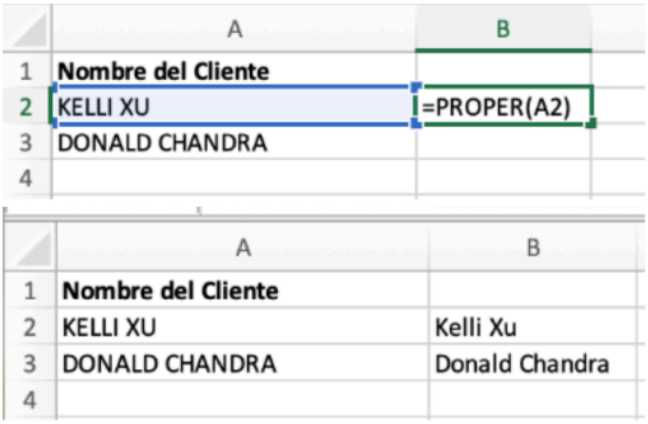
Esto es especialmente importante para los nombres propios. Más adelante, aprenderemos a comprobar y eliminar las entradas duplicadas. Es importante normalizar las mayúsculas primero, antes de eliminar los duplicados. Para entender por qué, imagine que hay entradas duplicadas para una persona llamada Kelli Xu, pero que las mayúsculas difieren (por ejemplo, "Kelli Xu" frente a "KELLI XU"), Excel no las tratará como valores duplicados hasta que las normalice.

Cambia la apariencia del texto utilizando los siguientes comandos:

- MAYUSC() - convierte el texto en todas las letras mayúsculas
- MINUSC() - convierte el texto en todas las letras minúsculas
- NOMPROPIO() - convierte el texto para que la primera letra de cada palabra sea mayúscula; el resto, minúscula

Nota: Para los nombres, queremos utilizar las mayúsculas "adecuadas" (utilice la función "NOMPROPIO()"). Rellene este comando hacia abajo de la columna utilizando la técnica descrita anteriormente para asignar valores de indicadores de resultado.

Figura 4. Cambio de los valores en mayúsculas a las mayúsculas adecuadas



Concepto 3. Eliminar / sustituir caracteres

Paso 3a. Eliminar los espacios extra y los caracteres no impresos. También es importante realizar este paso antes de eliminar los duplicados. Imagínese que el valor duplicado de Kelli Xu contiene un espacio erróneo al final de su nombre (por ejemplo, "Kelli Xu" frente a "Kelli Xu"), Excel no los tratará como valores duplicados.

Paso 3b. Sustituya los caracteres con acento por el carácter sin acento. Por ejemplo, Excel no tratará "María Ramírez" como un valor duplicado de "María Ramírez".

Identificar, eliminar y reemplazar caracteres utilizando los siguientes comandos / funciones:

- Utilizar la función "ESPACIOS()" para eliminar los espacios extraños del texto.
- Utilizar "**Buscar y reemplazar**" para identificar caracteres específicos y reemplazarlos por otros caracteres (por ejemplo, encontrar la "ñ" y reemplazarla por la "n").

Concepto 4. Comprobar los valores que faltan

A menudo hay valores que faltan en los conjuntos de datos. A veces, estos valores faltan para variables que son menos importantes para nuestro análisis. Por ejemplo, no pasa nada si no tenemos los segundos nombres de algunos de los destinatarios de nuestros correos electrónicos, porque conocer sus segundos nombres no afecta a nuestra comprensión del

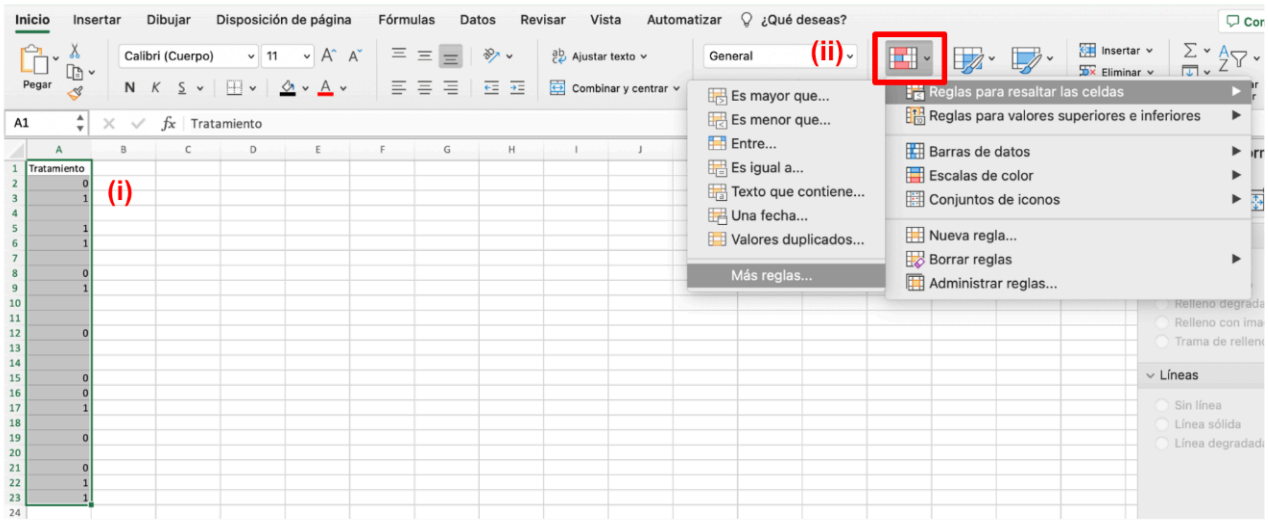
efecto de los correos electrónicos sobre los destinatarios que hacen clic en un enlace. Por otro lado, si no estamos seguros de qué versión de nuestro correo electrónico recibió un grupo de destinatarios porque faltan datos de la asignación del tratamiento, esto complica nuestra capacidad para identificar la versión más efectiva.

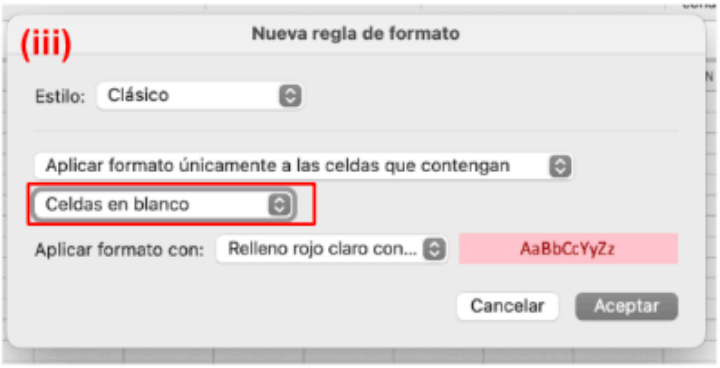
Una buena forma de comprobar los valores que faltan es utilizar el formato condicional para resaltarlos.

Para resaltar los valores que faltan (Ver Figura 5.)

i. Seleccione el rango de valores que desea comprobar. Puede tratarse de una variable, de varias variables o de todo el conjunto de datos. *No seleccione toda la(s) columna(s), ya que de lo contrario se resaltarán todas las celdas vacías de su conjunto de datos.*
ii. En la pestaña Inicio, selecciona Formato condicional >> Reglas de celdas resaltadas >> Más reglas...
iii. En la casilla "Formatear sólo las celdas que contienen", seleccione "Espacios en blanco", y haga clic en Aceptar. Todas las celdas vacías deberían quedar resaltadas.

Figura 5. Formato condicional para los valores perdidos





	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 ordenar los valores perdidos *(es decir, llevar las observaciones con valores perdidos a la parte superior del conjunto de datos)*

i. Seleccione todos los valores de su conjunto de datos.
ii. En la pestaña Inicio, seleccione Ordenar y filtrar >> Filtrar.
iii. Haga clic en la flecha hacia abajo en la etiqueta de la columna de interés.
iv. En el menú "Por color" en Ordenar, elija Color de la celda y el formato de la celda resaltada.

Para filtrar los valores perdidos (es decir, ver solo las observaciones con valores perdidos)
Nota: cuando se filtran las observaciones, éstas se ocultan, no se eliminan.

i. Seleccione todos los valores de su conjunto de datos.
ii. En la pestaña Inicio, seleccione Ordenar y filtrar >> Filtrar.
iii. Haga clic en la flecha hacia abajo en la etiqueta de la columna de interés.
iv. En el menú "Por color" en Filtro, elija Color de celda y el formato de celda resaltado. Sólo debería ver las observaciones con valores perdidos en esta columna.

Nota: Para borrar el formato condicional, en la pestaña Inicio, seleccione Formato condicional >> Borrar reglas >> Borrar reglas de toda la hoja o Borrar reglas de las celdas seleccionadas.

Concepto 5. Comprobar si hay observaciones duplicadas

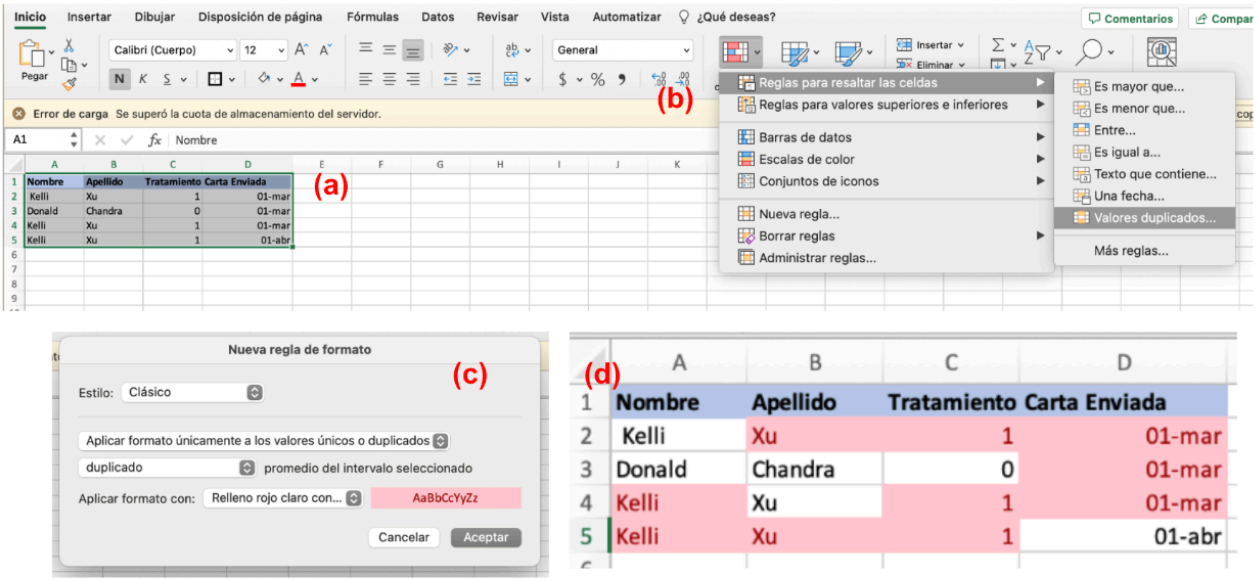
Compruebe los nombres y las direcciones para ver si hay duplicados. Si hay filas duplicadas completamente idénticas, elimínelas.

Por ejemplo, tomemos un conjunto de datos en el que cada observación es una persona, y tenemos su nombre completo. Habrá muchos nombres duplicados y puede que haya apellidos duplicados, así que normalmente se trata de identificar si el nombre completo es o no un

duplicado. Si el conjunto de datos ya incluye una columna con el nombre completo de una persona, elija esta variable cuando encuentre o elimine duplicados. Si una columna incluye el nombre de una persona y otra columna incluye su apellido, elija ambas columnas cuando busque duplicados. (Para los pasos siguientes, véase la figura 6).

Paso 5a. Seleccione sus datos.
Paso 5b. En la pestaña Inicio, seleccione Formato condicional >> Resaltar reglas de celdas >> Duplicar valores...
Paso 5c. En la ventana Nueva regla de formato, haga clic en Aceptar. (Puede cambiar el color del resaltado en el menú desplegable "Formato con".
Paso 5d. Los valores duplicados serán ahora resaltados.

Figura 6. Comprobación de valores duplicados



(!) En el Concepto 6, aprenderá a eliminar las observaciones duplicadas, pero pensemos si las observaciones duplicadas tienen sentido o si debemos eliminarlas.

En el ejemplo anterior, vemos que hay observaciones idénticas para Kelli Xu. En un escenario en el que nuestra intervención implique el envío de una carta por persona, podríamos llegar a la conclusión de que deberíamos eliminar la observación duplicada de Kelli Xu. No querríamos contar sus resultados dos veces al analizar los datos.

Sin embargo, imaginemos una intervención en la que un grupo de participantes recibe una carta, mientras el otro grupo recibe la misma carta *ADEMÁS DE* una carta de recordatorio como seguimiento. Después de resaltar los valores duplicados, nuestro conjunto de datos podría tener un aspecto similar al de la Figura 7 que aparece a continuación.

Querríamos eliminar la fila 4, que es un duplicado exacto de la fila 2, pero no querríamos eliminar la tercera observación de Kelli Xu en la fila 5, ya que esta línea muestra cuándo se envió su segunda carta.

Figura 7. Comprobación de valores duplicados, continuación

	A	B	C	D
1	**Nombre**	**Apellido**	**Tratamiento**	**Carta Enviada**
2	Kelli	Xu	1	01-mar
3	Donald	Chandra	0	01-mar
4	Kelli	Xu	1	01-mar
5	Kelli	Xu	1	01-abr
6				

La clave es pensar en su intervención y luego preguntarse: "¿Tiene sentido que haya valores duplicados?".

Concepto 6. Eliminar observaciones duplicadas

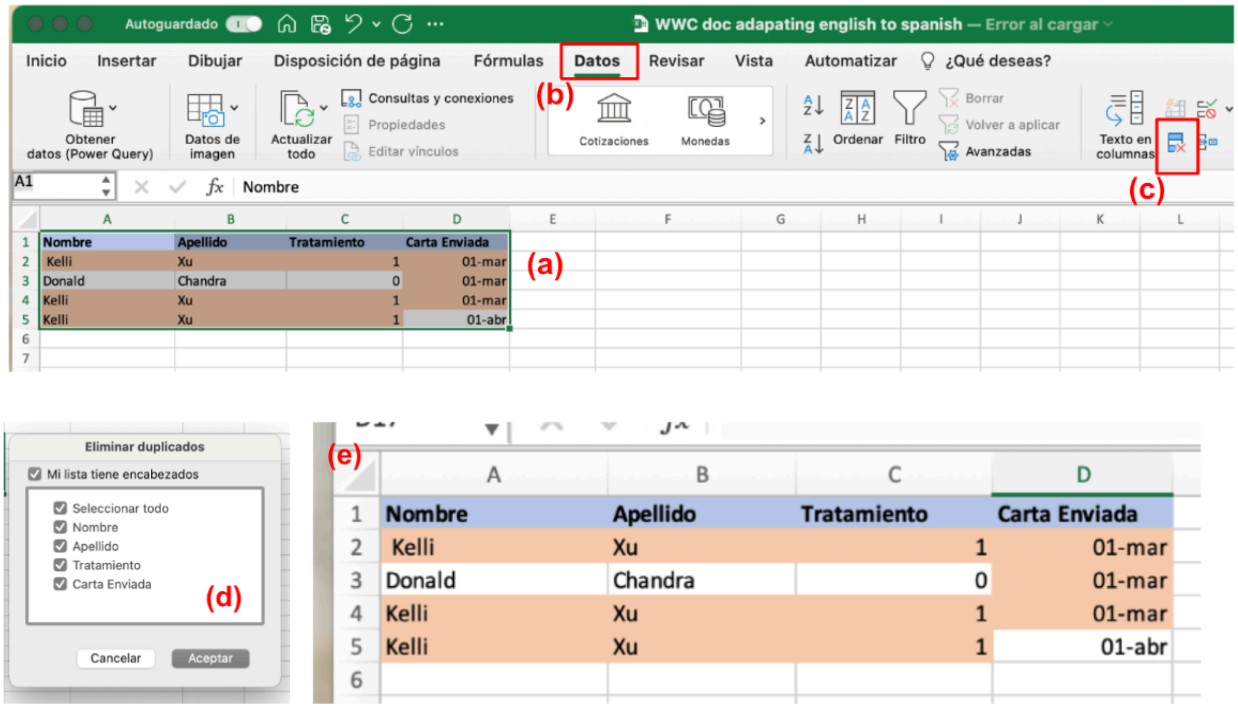
Si cree que una observación duplicada debe ser eliminada, siga los siguientes pasos para eliminarla. (Véase la figura 8.)

Paso 6a. Seleccione sus datos.
Paso 6b. Vaya a la pestaña Datos
Paso 6c. Haga clic en el botón Eliminar Duplicados.
Paso 6d. En la ventana de Eliminar Duplicados, asegúrese de que la opción "Mi lista tiene encabezados" está marcada si ha seleccionado todos los datos, incluyendo los encabezados. Luego asegúrese de que todas las columnas estén seleccionadas. Haga clic en Aceptar.
Paso 6e. Las observaciones duplicadas serán ahora eliminadas.

(!) Observe que solo se ha eliminado la entrada duplicada de Kelli Xu que recibió la carta el 1 de marzo, pero no la entrada que indica que Kelli Xu recibió la carta el 1 de abril. Esto se debe a que hemos seleccionado todas las columnas en la ventana Eliminar duplicados. Seleccionar todas las columnas significa que la observación tendría que ser un duplicado en todas estas columnas.

Si no hubiéramos seleccionado la columna D, Excel habría comprobado las observaciones que son idénticas para las columnas A, B y C. En ese caso, se habrían eliminado las filas 4 y 5.

Figura 8. Eliminación de entradas duplicadas



Algunas otras funciones que pueden ser útiles.
Las cadenas son simplemente piezas de texto que pueden ser manipuladas, consultadas, movidas y editadas utilizando funciones adicionales estándar de Excel. Para los ejemplos de las siguientes funciones, haremos referencia a los datos de la imagen de abajo.

	A	B
1	**Nombre**	**Apellido**
2	Kelli	Xu

- CONCATENAR() – une varias cadenas
  - Ex: =CONCATENAR(A2," ",B2) produce "Kelli Xu".
- IZQUIERDA() – devuelve los *n* caracteres izquierdos de una cadena
  - Ex: =IZQUIERDA(A2,3) produce "Kel".
- DERECHA() – devuelve los *n* caracteres de la derecha de una cadena
  - Ex: =DERECHA(A2,2) produce "li".
- EXTRAE() – devuelve los caracteres de la mitad de una cadena
  - Ex: =EXTRAE(A2,2,3) produce "ell".

Análisis de Datos

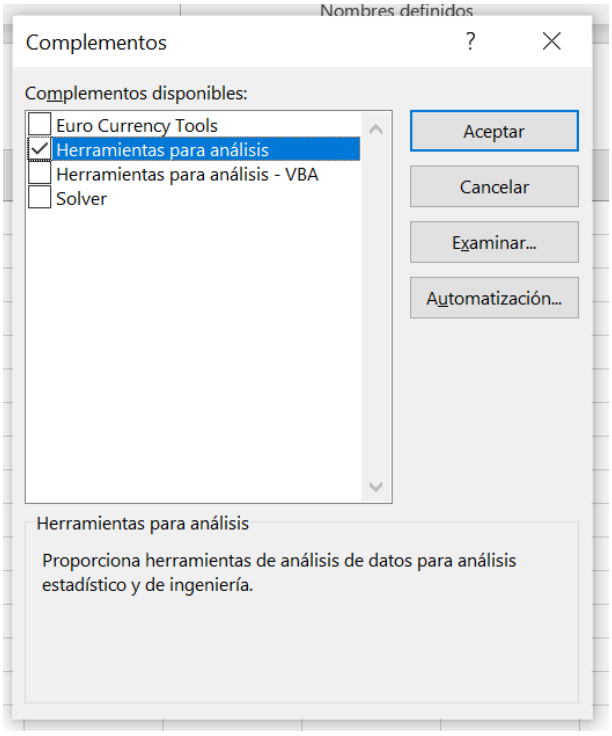
Después de limpiar los datos, está listo/a para empezar a analizarlos. El análisis de datos nos ayuda a comprender mejor nuestros datos, descubrir patrones, sacar conclusiones y fundamentar la toma de decisiones. Esta sección proporciona una visión general básica de cómo describir los datos y sacar conclusiones de un análisis de regresión básico de un ensayo controlado aleatorizado.

¿Qué vas a aprender a hacer?

1. Elaborar e interpretar estadísticas descriptivas
2. Realizar una regresión
3. Interpretar los resultados de la regresión
4. Representar gráficamente los resultados de la regresión

Para realizar el análisis estadístico, utilizaremos el complemento Analysis Toolpak de Excel (Instrucciones de instalación).

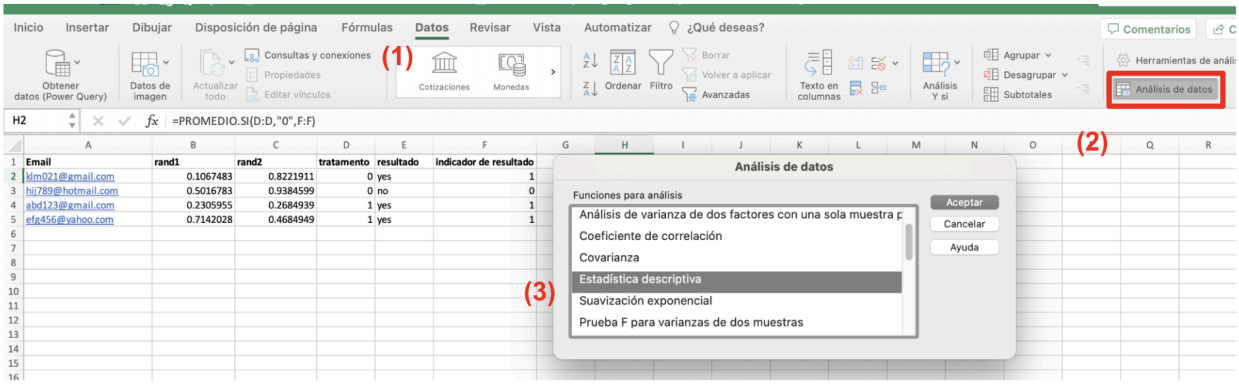
Figura 9. Complemento del paquete de herramientas de análisis de Microsoft Excel



Paso 1. Elaborar e interpretar las estadísticas descriptivas

Tras la instalación, en la pestaña "Datos" (1), utilice la función "Análisis de datos" (2) y seleccione "Estadísticas descriptivas" (3). Haga clic en "Aceptar". (Véase la Figura 10.)

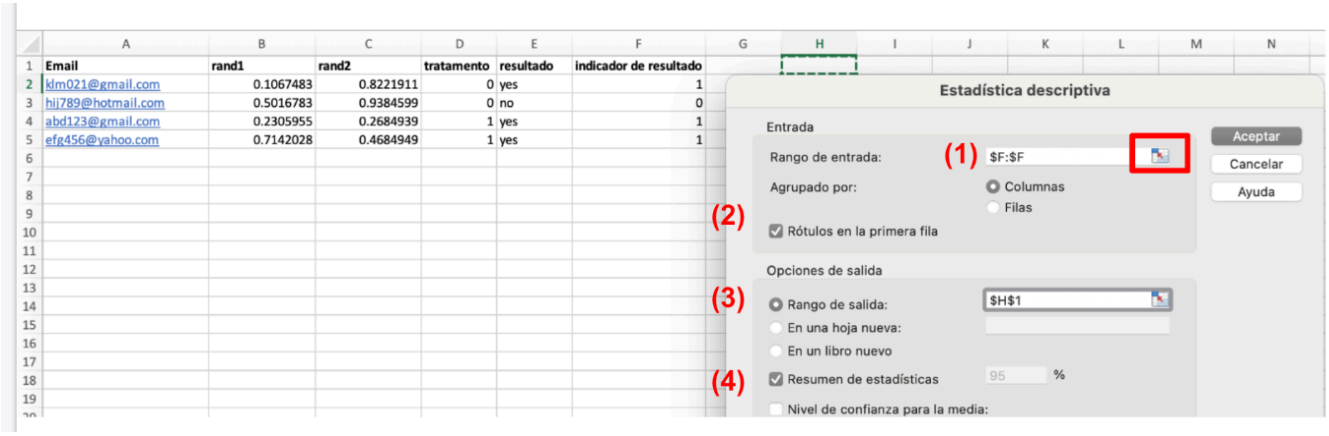
Figura 10. Acceso a la herramienta de estadísticas descriptivas



En la ventana "Estadísticas descriptivas", seleccione los datos de la variable para la que desea producir estadísticas descriptivas como "Rango de entrada" haciendo clic en el botón (1). Puede a) elegir toda la columna (en este caso, haciendo clic en la columna F) y marcar "Etiquetas en la primera fila" (2) o b) seleccionar sólo los valores bajo la etiqueta de la variable y asegurarse de que "Etiquetas en la primera fila" no esté marcada.

A continuación, elija dónde desea que aparezca la tabla de estadísticas de resumen, haciendo clic en el botón situado junto a "Rango de salida" (3). Puede elegir donde quiera que aparezca la tabla, y la celda que seleccione será la esquina superior izquierda de la tabla. En este ejemplo, hemos seleccionado H1 como inicio de la tabla. Por último, marque la opción "Estadísticas de resumen" (4) y haga clic en "Aceptar". (Véase la Figura 11.)

Figura 11. Elaboración de las estadísticas descriptivas



Después de hacer clic en "Aceptar", aparecerá la tabla de estadísticas de resumen para esta variable donde se establece el rango de salida y tendrá un aspecto similar al de las dos primeras columnas de la tabla de la Figura 12. En la tercera columna, hemos incluido descripciones para conceptos comunes.

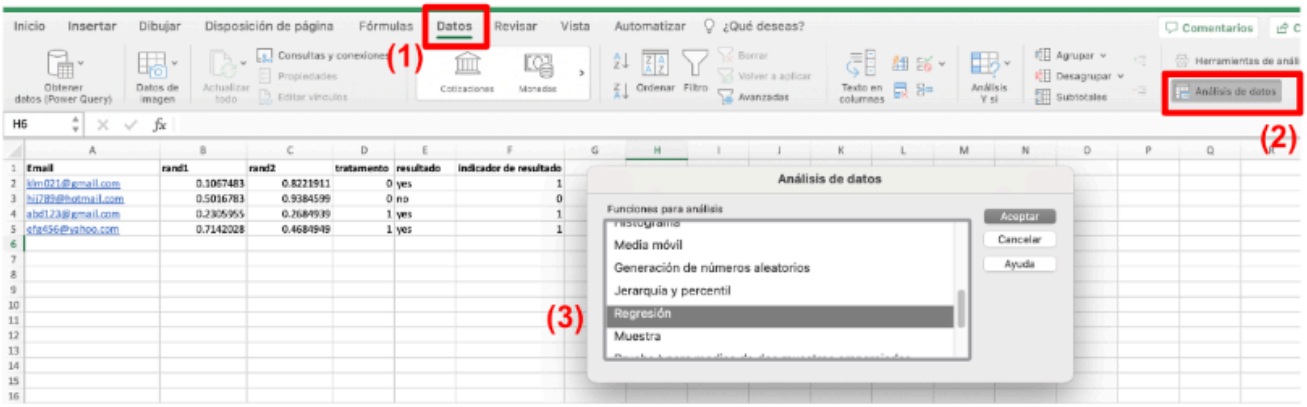
Figura 12. Tabla de resumen estadístico con descripciones

resultado		Descripción
Media	0.75	Media de los valores
Error estándar	0.25	Medida que indica con qué precisión se aproxima la media muestral a la media poblacional. Cuanto mayor sea el número, menos probable es que la media muestral sea exacta
Mediana	1	Punto medio de los valores
Moda	1	Valor que se repite con más frecuencia
Desviación estándar	0.5	La cantidad de variación de un conjunto de valores. Es decir, lo similares que son los resultados entre las unidades de la muestra. Pongamos un ejemplo en el que la duración media del tiempo que tarda alguien en pagar una multa es de 60 días. Esta media podría tener una desviación típica pequeña si el -68% de las personas pagan su multa entre 50 y 70 días, o una desviación típica grande si el -68% de las personas pagan su multa entre 10 y 110 días. Una desviación típica mayor indica una mayor "dispersión".
Varianza de la muestra	0.25	
Curtosis	4	
Sesgo	-2	
Rango	1	Diferencia entre el valor más alto y el más bajo
Mínimo	0	Valor más bajo
Máximo	1	Valor más alto
Suma	3	Suma de todos los valores
Recuento	4	Número de valores

Paso 2. Realizar una regresión

En la pestaña "Datos" (1), utilice la función "Análisis de datos" (2) y seleccione "Regresión" (3) como herramienta de regresión. Haga clic en "Aceptar". (Véase la Figura 13.)

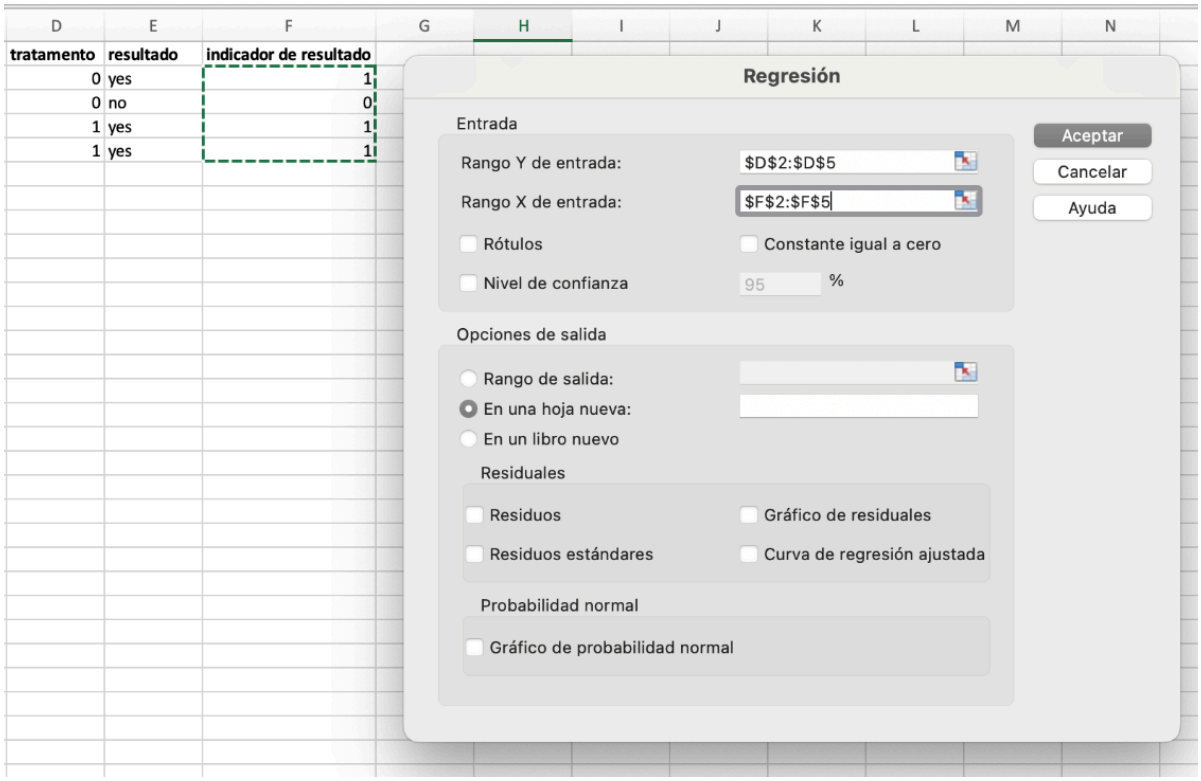
Figura 13. Acceso a la herramienta de regresión



En la configuración de la regresión, introduzca la columna del "indicador de resultado" como "Rango Y" y la columna del tratamiento como "Rango X". Puede a) elegir toda la columna y

marcar "Etiquetas" o b) seleccionar sólo los valores bajo la etiqueta de la variable y asegurarse de que "Etiquetas" no esté marcada (véase la figura 14).

Figura 14. Producción de la regresión



Una vez que pulsemos ACEPTAR, el programa creará una nueva hoja de trabajo con la salida de la regresión similar a la de la Figura 15.

Figura 15. Resultado de la regresión

	A	B	C	D	E	F	G	H	I
1	Resumen de resultados								
2									
3	*Estadísticas de Regresión*								
4	Coeficiente de correlación múltiple	0.1389467							
5	R al cuadrado	0.0164784							
6	R al cuadrado ajustada	-0.0184383							
7	Error estándar	0.5968572							
8	Observaciones	30							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Siginificancia F*			
12	Regresión	1	0.1333333	0.13333	0.5283	0.47339			
13	Residual	28	7.0666667	0.25238					
14	Total	29	7.2						
15									
16		Coeficientes	Error Estándar	t-Stat	Valor-P	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
17	Intercepción	0.3333333	0.1297853	2.56893	0.01579	0.06783	0.569392	0.06739	0.59904
18	X Variable 1	0.13333333	0.18344853	0.72684	0.47895	-0.45256	0.57393	-0.24352	0.5091
19									

Paso 3. Interpretar los resultados de la regresión

La tabla de regresión contiene bastante información, pero hay dos lugares importantes en los que hay que fijarse 1) en el número de observaciones (identificado en la fila inferior de la primera tabla) y 2) en la tabla inferior.

En primer lugar, debe fijarse en el número de observaciones. ¿Es el que esperaba? ¿Coincide con el número de filas de sus datos?

A continuación, observe la tabla inferior, que contiene sus resultados clave. La tabla debe tener 2 filas: "Intercepción" y "Variable X". La fila "Intercepción" contiene información sobre el **primer grupo de correo electrónico (a menudo el grupo de control)** y la fila "Variable X" contiene información sobre el **segundo grupo de correo electrónico (a menudo el grupo de tratamiento)**. En nuestro escenario hipotético, **nos interesa comparar la proporción de individuos del primer grupo de correo electrónico que hacen clic en un enlace con la proporción de individuos del segundo grupo de correo electrónico**.

La tabla 1 incluye una descripción de cada estadística relevante.

Tabla 1. Descripción de las estadísticas de los resultados de la regresión

Estadístico	Variable X (Correo electrónico 2)
Coeficiente	La diferencia entre las medias del primer y segundo grupo de correo electrónico, a menudo denominada "efecto del tratamiento". Si vemos que el coeficiente de la "Variable X" es un número negativo, podemos interpretar que la media del segundo grupo es inferior a la del primero. Puede multiplicar el número por 100 para ver este valor en forma de porcentaje. En el ejemplo anterior, recibir el segundo correo electrónico aumentaría la probabilidad de que los participantes hagan clic en el enlace en un 13,33%.
Error estándar	Describe la probabilidad de que la media del segundo grupo de correos electrónicos de nuestra muestra refleje exactamente la media de la población. Es decir, si el 46,7% del segundo grupo de correos electrónicos hizo clic en el enlace, esta estadística nos indica la probabilidad de que el 46,7% de nuestra población de interés también haga clic en el enlace si recibe esta versión del correo electrónico.
Valor p	La probabilidad de identificar falsamente un efecto en sus datos que no existe en la realidad. La mayoría de los experimentos llaman a los resultados "estadísticamente significativos" cuando hay un 5% o menos de probabilidades de que el efecto que detectaron no exista en realidad.

Inferior 95%[1]	El límite inferior del intervalo de confianza del 95% del coeficiente para el segundo grupo de correos electrónicos. Recuerde que el coeficiente representa la diferencia entre el primer y el segundo grupo de correo electrónico. Sume el porcentaje de este valor al coeficiente de intercepción * 100. Esto le da el límite inferior del intervalo de confianza del 95% de la media para el segundo grupo de correos electrónicos.
Superior al 95%[2]	El límite superior del intervalo de confianza del 95% del **coeficiente** para el segundo grupo de correos electrónicos. Recuerde que el coeficiente representa la diferencia entre el primer y el segundo grupo de correo electrónico. Sume el porcentaje de este valor al coeficiente de intercepción * 100. Esto le da el límite superior del intervalo de confianza del 95% de la media para el segundo grupo de correos electrónicos.

Cálculo de la media del grupo de control

Si su regresión incluye covariables, tenemos que seguir unos sencillos pasos para calcular la media del resultado del grupo de control.

Para calcular la media, las variables "tratamiento" e "indicador de resultado" deben formatearse como una variable binaria con ceros y unos. (Véase el Concepto 1 en Limpieza de datos)

En su hoja de cálculo, introduzca el siguiente comando en cualquier celda vacía.

=PROMEDIO.SI(D:D,”0”,F:F)

[1] Un intervalo de confianza representa el rango en el que existe una probabilidad específica de que el valor de un parámetro determinado (en nuestro caso, la media de clics) para la población de interés se encuentre dentro de este rango. Una probabilidad estándar para un intervalo de confianza es el 95%. Por ejemplo, el intervalo de confianza del 95% para el primer grupo de correos electrónicos es del 6,7% al 59,9%. Esto significa que, si nuestra población de interés recibiera la primera versión del correo electrónico, existe una probabilidad del 95% de que la media de clics en el enlace se encuentre dentro de este intervalo. Los valores "Lower 95%" y "Upper 95%" representan los límites inferior y superior de este intervalo, respectivamente. Cuanto más pequeño sea el rango del intervalo de confianza, más seguros podemos estar de que el verdadero valor de nuestra variable de interés para la población está cerca del valor observado de la muestra.

[2] Ver la nota anterior.

H2 =PROMEDIO.SI(D:D,"0",F:F)

	A	B	C	D	E	F	G	H
1	**Email**	**rand1**	**rand2**	**tratamento**	**resultado**	**indicador de resultado**		
2	klm021@gmail.com	0.1067483	0.8221911	0	yes	1		0.5
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								

Este comando hace que Excel realice una prueba lógica: para todas las observaciones con un valor de "0" en la columna D (es decir, las observaciones del grupo de control), tome la media de los valores de la columna F.

En otras palabras, Excel hace tres cosas. (1) Excel cuenta el número de personas en el grupo de control. (2) Excel toma la suma de los resultados del grupo de control. Como hemos formateado esta variable como un cero o un uno, la suma acaba siendo el número de personas del grupo de control que hicieron clic en el enlace. (3) Excel toma la suma de los resultados del grupo de control y la divide por el número de personas del grupo de control.

Paso 4. Representar gráficamente los resultados de la regresión

Podemos visualizar este resultado en forma de gráfico utilizando la "Plantilla de Gráficos.xlsx" del BIT. Deberá introducir varios datos del resultado de la regresión (detallados anteriormente) en la plantilla de gráficos para crear un gráfico. También tendrá que introducir la media del grupo de control (detallada anteriormente). Las celdas a las que tendrá que hacer referencia desde el resultado de la regresión y las estadísticas descriptivas están resaltadas en la Figura 16 a continuación

Figura 16. Identificación de los valores resultantes de la regresión para la plantilla de gráficos

	A	B	C	D	E	F	G	H	I
3	*Estadísticas de Regresión*								
4	Coeficiente de correlación múltiple	0.1389467							
5	R al cuadrado	0.0164784							
6	R al cuadrado ajustada	-0.0184383							
7	Error estándar	0.5968572							
8	Observaciones	30							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Siginificancia F*			
12	Regresión	1	0.1333333	0.13333	0.5283	0.47339			
13	Residual	28	7.0666667	0.25238					
14	Total	29	7.2						
15									
16		Coeficientes	Error Estándar	t-Stat	Valor-P	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
17	Intercepción	0.3333333	0.1297853	2.56893	0.01579	0.06783	0.569392	0.06739	0.59904
18	X Variable 1	0.13333333	0.18344853	0.72684	0.47895	-0.45256	0.57393	-0.24352	0.5091
19									

Deberá introducir estos valores en las celdas que corresponden a cada uno de ellos en la plantilla de gráficos. La Figura 17 muestra una plantilla completa y la Tabla 2 muestra dónde encontrar cada valor en la salida de la regresión y dónde introducirlos en la plantilla de gráficos.

Figura 17. Rellenar la plantilla de gráficos para graficar los resultados de la regresión

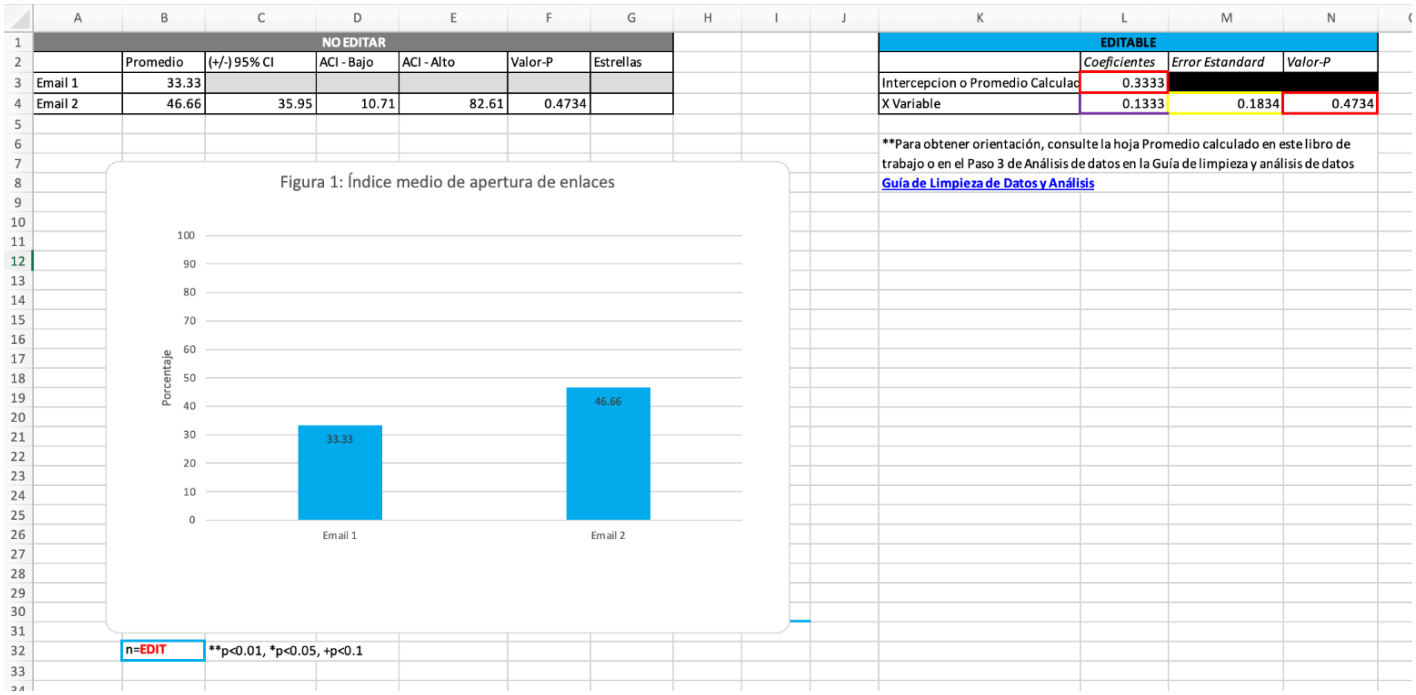


Tabla 2. Introducción de valores clave para producir el gráfico

Valor	Dónde puede encontrarlo (Resultado de la regresión)					Dónde escribirlo (Plantilla gráfica)			
Media del primer correo electrónico (o grupo de control)		*Coeficientes*	*Error Estándar*	*Estadístico T*	*Valor-P*	**EDITAR**			
	Intercepción	0.3333333	0.12971275	2.56978	0.01579		*Coeficientes*	*Error Estándar*	*Valor-P*
	Variable X 1	0.1333333	0.18344153	0.72684	0.47336	Intercepción	0.3333		
						Variable X	0.1333333	0.18344153	0.47336
Efecto del tratamiento		*Coeficientes*	*Error Estándar*	*Estadístico T*	*Valor-P*	**EDITAR**			
	Intercepción	0.3333333	0.12971275	2.56978	0.01579		*Coeficientes*	*Error Estándar*	*Valor-P*
	Variable X 1	0.1333333	0.18344153	0.72684	0.47336	Intercepción	0.3333		
						Variable X	0.1333333	0.18344153	0.47336
Error estándar		*Coeficientes*	*Error Estándar*	*Estadístico T*	*Valor-P*	**EDITAR**			
	Intercepción	0.3333333	0.12971275	2.56978	0.01579		*Coeficientes*	*Error Estándar*	*Valor-P*
	Variable X 1	0.1333333	0.18344153	0.72684	0.47336	Intercepción	0.3333		
						Variable X	0.1333333	0.18344153	0.47336
Valor P		*Coeficientes*	*Error Estándar*	*Estadístico T*	*Valor-P*	**EDITAR**			
	Intercepción	0.3333333	0.12971275	2.56978	0.01579		*Coeficientes*	*Error Estándar*	*Valor-P*
	Variable X 1	0.1333333	0.18344153	0.72684	0.47336	Intercepción	0.3333		
						Variable X	0.1333333	0.18344153	0.47336

Retoques finales
Una vez que hayamos terminado de rellenar estas casillas, podremos trabajar en los retoques finales:

- **Cambiar el título del gráfico.** Haga doble clic en el título del gráfico y escriba el nombre adecuado.
- **Ajuste el rango del eje vertical.** Dependiendo de los promedios de sus grupos y del intervalo de confianza, es posible que usted quiera ajustar el rango del eje vertical. Por ejemplo, si la media de un grupo es de 5,5% y la del otro es de 10,5%, sería difícil visualizar la diferencia si el eje vertical se extiende de 0 a 100%. Puede ser más razonable tener un rango de 0 - 20%. (Véase la figura 18 para los pasos siguientes).
    1. Haga clic en el eje vertical.
    2. Debe aparecer la ventana de Formato de eje.
    3. En Opciones del eje, cambie el límite máximo. (*Para el ejemplo anterior, este valor sería 20*)
    4. Si lo desea, cambie la unidad mayor. La unidad mayor representa el espacio entre cada marca de verificación en el eje vertical. En esta plantilla, la unidad mayor se establece como 10. (*Para el ejemplo anterior, podría tener sentido cambiar esto a 4 o 5*)
- **Actualice el tamaño de la muestra.** Haga doble clic donde dice "n = EDIT" en la nota al pie. Sustituya "EDIT" por el valor de la celda de "Observaciones" de la tabla de salida, que aparece en azul. Asegúrese de cambiar el color de la fuente del tamaño de la muestra a negro. (Véase la Figura 19 a continuación).

Figura 18. Ajuste del rango del eje vertical

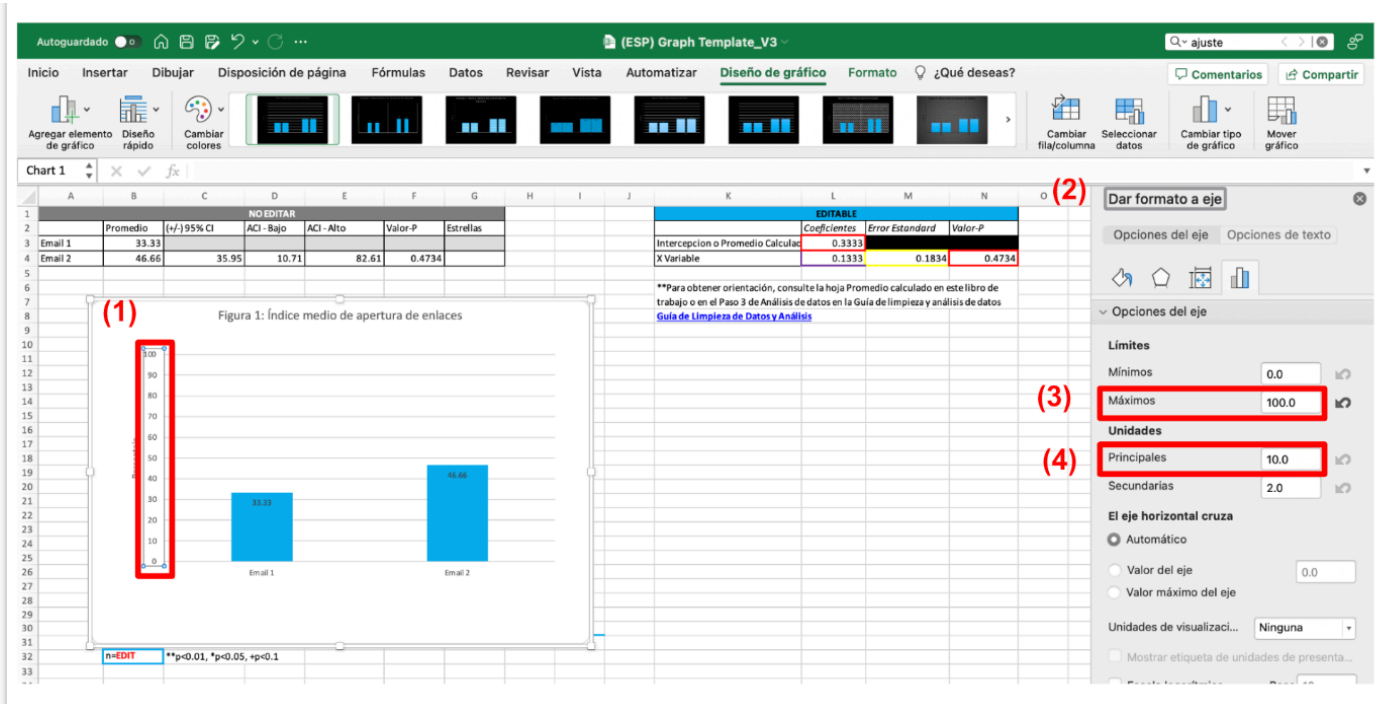
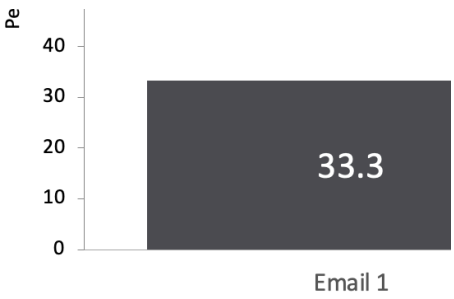


Figura 19. Actualización del tamaño de la muestra



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

8		
9	*Estadísticas de Regresión*	
10	Coeficiente de correlación múltiple	0.1389467
11	R al cuadrado	0.0164784
12	R al cuadrado ajustada	-0.0184383
13	Error estándar	0.5968572
14	Observaciones	30
15		

El gráfico está ahora totalmente listo para ilustrar el resultado de la evaluación en el informe.