Laboratorio 3 - Preparación de Datos

Objetivos

En esta práctica de laboratorio, aprenderá los aspectos básicos de la limpieza de datos en Microsoft Excel. Los datos deben limpiarse antes de poder analizarse. La limpieza de los datos facilita su lectura e interpretación, garantiza la uniformidad y resultados precisos, y permite un mejor proceso de toma de decisiones.

Parte 1: Explorar un Conjunto de Datos de Muestra

Parte 2: Limpieza de Datos

Trasfondo / Escenario

El rápido aumento del volumen y la complejidad de los datos se deben al creciente tráfico de datos móviles, el tráfico de computación en la nube y la adopción de tecnologías como la Internet de las Cosas y la Inteligencia Artificial. Los datos en su forma bruta son inútiles a menos que puedan limpiarse y analizarse para proporcionar información que conduzca al conocimiento y la sabiduría para una inteligencia procesable. Las empresas están descubriendo rápidamente que en realidad no quieren los datos en sí: quieren la información y el conocimiento que proviene de los datos y que pueden utilizar para tomar mejores decisiones.

Recursos necesarios

- Dispositivo móvil o PC/computadora portátil con conexión a Internet y Microsoft Excel o un programa de hoja de cálculo similar

Nota: Los pasos precisos para formatear y manipular datos en Excel pueden variar entre plataformas, lenguajes y versiones. Las instrucciones de esta práctica de laboratorio se basan en la versión gratuita de Excel disponible en Office.com y es posible que deban modificarse para que coincidan con la plataforma, el software, el lenguaje o la versión del usuario a fin de lograr los resultados que se muestran en esta práctica de laboratorio.

Instrucciones

Parte 1: Explorar un Conjunto de Datos de Muestra

Paso 1: Abra Bike Sales_Prepare_Lab 3.4.7.xlsx

1. Descargue el archivo de ejemplo de Excel titulado **Bike Sales_Prepare_Lab 3.4.7.xlsx** y explórelo.
   Una variedad de errores de datos se incluyen intencionalmente en la hoja de trabajo de

Venta de Bicicletas para este ejercicio. Las herramientas utilizadas en esta práctica de laboratorio le permitirán encontrar y corregir estos errores.

Paso 2: Expanda las Columnas de la Hoja de Datos según sea necesario

1. Notará que algunos de los datos están truncados en las columnas. Expanda las columnas para poder revisar los datos.

Paso 3: Revise los Datos.

1. Revise los datos sin procesar para localizar cualquier dato que pueda sesgar el análisis de datos.

Enumere cualquier problema en los datos que pueda afectar su análisis.

> Por ejemplo: celdas en blanco, datos que deben dividirse en columnas y ceros para la Cantidad Pedida.

Parte 2: Limpieza de datos

Paso 1: Búsqueda de duplicados

El conjunto de datos proporcionado puede contener entradas duplicadas. Uno de los procesos de limpieza de datos es encontrarlos y eliminarlos. En el conjunto de datos de ventas de bicicletas, la única columna que no puede tener el mismo valor más de una vez es la columna A, N°_Pedido_Venta (Sales_Order #).

1. Seleccione la columna A para verificar si hay datos duplicados.
2. Con la columna A seleccionada, haga clic en el botón **Formato Condicional** (Conditional Formatting) en la barra de herramientas **Inicio** (Home) y seleccione **Resaltar Reglas de Celda** (Highlight Cell Rules) > **Valores Duplicados** (Duplicate Values) y, a continuación, haga clic en **Listo** (Done).
   El formato condicional debe encontrar dos pares de duplicados. Las celdas A2 y A3 tienen el mismo número de pedido de venta 261695. Además, las celdas A8 y A9 tienen el mismo número de pedido de venta 261701. Estos tipos de entradas duplicadas pueden ocurrir fácilmente durante la entrada manual de datos o al copiar y pegar datos en una hoja de trabajo.

Revise las entradas duplicadas.

En el caso de las celdas **A2** y **A3**, parece que el número de pedido de venta 261695 se ingresó incorrectamente en la celda **A3**. Como analista de datos, deberá ir a la fuente de los datos y verificar el número de pedido de venta.

Para las celdas **A8** y **A9**, una revisión detallada muestra que ambas filas son exactamente iguales. Lo más probable es que en este caso se haya ingresado dos veces una entrada de ventas.

Paso 2: Reparación y Eliminación de Duplicados

Cuando se identifican entradas de datos duplicadas, deben revisarse cuidadosamente antes de eliminarse, para que los datos relevantes no se eliminen accidentalmente.

1. Para corregir la entrada duplicada en la celda **A3**, cambie el N.º de Pedido de Venta a **261696**. Excel debe eliminar automáticamente el formato condicional.
2. Para corregir las filas duplicadas 8 y 9, se debe eliminar una de las filas.
    1. Una forma de eliminar una entrada duplicada es seleccionar una fila y eliminarla.
    2. Si hay muchos duplicados que deben eliminarse en un conjunto de datos grande, se puede utilizar la herramienta Eliminar Duplicados (Remove Duplicates).
3. Haga clic en la herramienta **Eliminar Duplicados** (Remove Duplicates) en la barra de herramientas **Datos** (Data).
4. En el cuadro de diálogo **Eliminar Duplicados** (Remove Duplicates), seleccione la columna **N°_Pedido_Venta** (Sales_Order #) y asegúrese de que la casilla de verificación **Mis datos tienen encabezados** (My list has headers) esté seleccionada. Haga clic en **Aceptar** (OK) para continuar.
   Excel eliminará la segunda instancia de cada conjunto de filas duplicadas.

Paso 3: Encontrar celdas vacías

Hay muchas razones por las que una celda puede estar en blanco. Podría ser un error humano debido al ingreso manual de datos o podría ser el resultado de la copia de datos de otras fuentes. El contexto es clave al determinar qué hacer con las celdas vacías. A veces, un analista de datos deberá completar cada celda en blanco de los datos con el mismo valor constante. Otras veces, puede haber pistas de los datos circundantes sobre lo que debería estar en una celda vacía. Es posible que los analistas también tengan que volver a la fuente de los datos para descubrir cuáles deberían ser los valores faltantes.

Para buscar celdas vacías, se puede utilizar la herramienta **Formato Condicional**.

1. Seleccione la hoja completa haciendo clic en la flecha en la esquina superior izquierda de la hoja de trabajo a la izquierda de la columna A.
2. Haga clic en la herramienta **Formato Condicional** (Conditional Formatting) en la barra de herramientas **Estilos** (Styles) y seleccione **Resaltar Reglas de Celda** (Highlight Cell Rules) > **Texto que Contiene** (Text That Contains).
3. En la ventana **Formato Condicional** (Conditional Formatting), en **Tipo de Regla** (Rule Type), seleccione **Espacios en Blanco** (Blanks). Cambie **Formato** (Format) a **Relleno verde con texto verde oscuro** (Green fill with dark green text) y haga clic en **Listo** (Done).
   Cualquier celda en blanco en la hoja de trabajo ahora debe llenarse en verde. Debe haber cuatro celdas resaltadas: C12, G17, M23 y N24.
   Es probable que un analista de datos revise las celdas en blanco para ver si se pueden obtener los datos que faltan. Si no puede, la única opción puede ser eliminar las filas de los datos faltantes.

4. Para simular haber encontrado los datos faltantes, complete las celdas vacías de la siguiente manera:
   1) **C11 = 5**
   2) **G16 = Youth (<25)**
   3) **M22 = Mountain-200 Black, 42**
   4) **N23 = 4**
   Paso 4: Análisis de Datos de Texto a Columna

Observará que algunas de las celdas tienen varios elementos de datos separados por un delimitador de datos como una coma. Por ejemplo, considere la columna **Descripcion_Producto** (Product_Description). Puede analizar los datos en esta columna para que cada parte de la descripción del producto se muestre en su propia columna. Utilizará la función **Texto a Columnas** para lograr esto.

Analizará los datos en la columna **Descripción_Producto** para mover el tamaño y el color de la bicicleta en columnas separadas.

1. Comience agregando una nueva columna en blanco a la derecha de la columna M, **Descripción_Producto**. Esta nueva columna se convierte en la columna N.
2. Resalte la columna M, **Descripción_Producto**.
3. En la barra de herramientas **Datos** (Data), haga clic en el botón **Texto a columnas** (Text to Columns) en la cinta de opciones **Herramientas de Datos** (Data Tools).
4. En la ventana **Texto a Columnas** (Text to Columns), seleccione **Coma** (Comma) como único delimitador y haga clic en **Aplicar** (Apply).
5. Todos los tamaños de bicicletas ahora deben moverse a la columna N.
6. Ponga el nombre a la columna N “**Tamaño** (Size)”.
   Luego, repita este proceso para eliminar el color de la bicicleta de la columna **Descripción_Producto**.
7. Agregue una nueva columna nuevamente a la derecha de la columna M.
8. Use la herramienta **Texto a Columnas** (Text to Columns) nuevamente, pero esta vez establezca el delimitador en **Espacio** (Space).
   Esto mueve el color de las bicicletas a la nueva columna para que todo lo que quede en la columna M ahora sean los modelos de las bicicletas.
9. Nombre la nueva columna “**Color**”.
10. Cambie el nombre de la columna M de Product_Desciption a **Model** (Modelo).
11. Ahora debería tener tres columnas, cada una de las cuales muestra una parte de la descripción de la bicicleta.

M	N	O
Model	**Color**	**Size**
Mountain-200	Black	46
Mountain-200	Silver	42
Mountain-400-W	Silver	46

Paso 5: Eliminación de Espacios Adicionales

Cuando se pegan datos de fuentes externas en una hoja de cálculo de Excel, existe una buena probabilidad de que las celdas contengan espacios adicionales que deban eliminarse para que las búsquedas y las consultas generen resultados precisos. La función ESPACIOS o TRIM se utiliza para eliminar el exceso de espacios y tabulaciones en las celdas de la hoja de cálculo de Excel. En este paso, utilizará las funciones de ESPACIOS o TRIM para limpiar los datos.

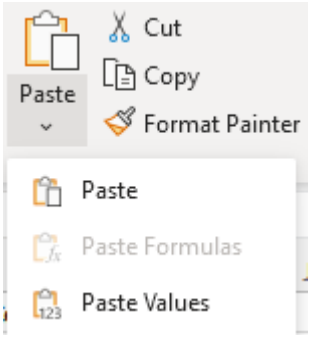
1. Inserte dos nuevas columnas en blanco a la derecha de la columna I, **País**. Se convertirán en las columnas J y K.
2. Nombre las dos nuevas columnas Longitud (Length) y ESPACIOS (TRIM), respectivamente.
3. En la columna Longitud, en la celda **J2**, ingrese la función **=LARGO(I2) o =LEN(I2)** para ver cuántos caracteres hay en la celda **I2**.
   El resultado debe ser 13, que es la cantidad de caracteres en “Estados Unidos” (United States) si se cuenta el espacio en el medio.
4. Ahora copie esta función **LARGO o LEN** de la celda **J2** hasta la celda **J8**.
   Observe que las otras celdas con Estados Unidos (United States) muestran una longitud de 14 caracteres y no 13. Esto se debe a que cada una de estas celdas contiene un espacio adicional.
   En la celda **I4**, el espacio adicional está al frente. En la celda **I6**, el espacio adicional está entre las palabras United y States. En la celda **I8**, el espacio adicional no se nota fácilmente, pero está al final.

	I	J
1	**Country**	**Length**
2	United States	13
3	United Kingdom	14
4	United States	14
5	Australia	9
6	United States	14
7	United Kingdom	14
8	United States	14

Para eliminar los espacios adicionales en estas celdas, use la función ESPACIOS o TRIM en la columna K.

5. En la celda **K4**, ingrese **=ESPACIOS(I4) o =TRIM(I4)**. La función elimina el espacio inicial de la celda **I4**.
6. Copie la función ESPACIOS (TRIM) en las celdas **K6** y **K8**. La función elimina los espacios de estas celdas.
7. Para asegurarse de eliminar los espacios adicionales en la columna País, copie y pegue la función **ESPACIOS** (TRIM) en todas las celdas de la columna K.
   Se eliminan todos los espacios adicionales en la columna M. Los valores de la columna ESPACIOS ahora deben pegarse en la columna País (Country).

8. Seleccione las celdas **K2** a **K89** y cópielas.
9. Seleccione las celdas **I2** a **I89** y haga clic en la flecha hacia abajo debajo de la herramienta **Pegar** (Paste) y, a continuación, seleccione **Pegar Valores** (Paste Values).



Esto pega solo los valores (nombres de países) en la columna I, y no las fórmulas que estaban en la columna K.

10. Elimine las columnas Longitud y ESPACIOS, ya que ya no son necesarias.

Paso 6: Cambio de mayúsculas y minúsculas

Al importar datos a Excel desde otra fuente, los datos a menudo se muestran en mayúsculas o minúsculas.

Excel le permite cambiar el texto a mayúsculas, minúsculas o mayúsculas. Las funciones MAYUSC (UPPER), MINUSC (LOWER), NOMPROPIO (PROPER) le permiten cambiar el uso de mayúsculas y minúsculas en el texto.

=MAYUSC(dirección de celda) - para conversión a mayúsculas

=MINUSC(dirección de celda) - para conversión a minúsculas

=NOMPROPIO(dirección de celda): para la conversión a mayúsculas en la primera letra de cada palabra.

En este ejemplo, los nombres de los países se cambiarán a minúsculas.

1. Agregue una nueva columna a la derecha de la columna I, País (Country). Esta nueva columna se convertirá en la columna J.
2. Etiquete la nueva columna como MINUSC (LOWER).
3. En la celda **J2**, ingrese la función **= MINUSC(I2) o =LOWER(I2)**
   El resultado debe ser “united states” (estados unidos), todo en minúsculas.
4. Copie y pegue la fórmula en las celdas **I3** a **I89** para poner todos los nombres de países en minúsculas.

1	Country	LOWER
2	United States	united states
3	United Kingdom	united kingdom
4	United States	united states
5	Australia	australia

5. Copie todos los nombres de países en la columna MINUSC (LOWER) columna J, y utilice **Pegar** > **Pegar valores** para pegarlos en la columna País, columna I.
6. Experimente con las funciones MAYUSC (UPPER) y NOMPROPIO (PROPER) de la misma manera.
7. Una vez finalizado, elimine la columna LOWER, ya que ya no es necesaria.

Paso 7: Resalte los Posibles Errores

En este paso, resaltará todos los costos unitarios y precios unitarios que sean cero. Este tipo de datos es falso y sesgará el conjunto de datos, por lo que es importante encontrar estos errores y corregirlos.

1. Seleccione la columna Costo_Unitario (Unit_Cost) y Precio_Unitario (Unit_Price).
2. Haga clic en la herramienta **Formato Condicional** (Conditional Format) > **Resaltar Reglas de Celda** (Highlight Cell Rules) > **Igual a** (Equal To).
3. En la ventana Formato Condicional, introduzca **0** en el cuadro Igual a y haga clic en **Listo** (Done) para resaltar todas las bicicletas con un costo unitario o un precio unitario de cero.
   La celda **O9** en Costo_Unitario y la celda **P9** en Precio_Unitario deben resaltarse en rojo.
   Un analista de datos deberá determinar qué valores ingresar en estas celdas o eliminar estas dos filas de datos. En este ejemplo, conocemos los valores que deben estar en estas celdas, porque están en otras filas de la hoja.
4. Corrija el problema ingresando $1252 en la celda **O9** y $769 en la celda **P9**.

Paso 8: Buscar y Reemplazar

Buscar y reemplazar lo ayudará a ubicar y reemplazar datos en toda la hoja de trabajo.

En este ejemplo, supongamos que queremos reemplazar F y M en la columna Género_Cliente con las palabras completas Mujer y Hombre para facilitar la lectura.

1. Seleccione la columna **Género_Cliente** (Customer Gender).
2. Haga clic en el botón **Buscar y Seleccionar** (Find & Select) y seleccione **Reemplazar** (Replace).
3. En el cuadro de diálogo Buscar y Reemplazar, ingrese la siguiente información.
   1) **Buscar** (Find what): M
   2) **Reemplazar con** (Replace with): Hombre
   3) Expanda **Opciones de Búsqueda** (Options) y en **Dentro de** (Within): seleccione **Selección** (Selection)
4. Haga clic en **Buscar Todo** (Find All).
   Todas las celdas con un valor de M se resaltan.
5. Haga clic en **Reemplazar Todo** (Replace all).
   Deberían haberse reemplazado 38 coincidencias.

6. Para buscar y reemplazar “F” por “Mujer”, utilice el mismo proceso, excepto en **Opciones de Búsqueda** (Options) asegúrese de que **Dentro de** (Within): esté configurado en **Selección** (Selection) y que la opción **Distinguir Mayúsculas de Minúsculas** (Match case) esté activada. Esto evitará que Excel cambie la “f” minúscula si la hubiera, cambiando el sentido de otras palabras.

¿Cuántos reemplazos se realizaron de F a Mujer?

> 50 coincidencias.

Paso 9: Revisión Ortográfica

La función de verificación ortográfica está disponible en la pestaña **Revisar** (Review). Puede verificar la ortografía por celda, columna, fila u hoja. El corrector ortográfico garantiza que los errores ortográficos no provoquen que los resultados de las búsquedas o las consultas sean incorrectos.

1. Seleccione todas las columnas con valores de texto.
2. En la barra de herramientas **Revisar**, haga clic en la herramienta **Ortografía** (Spelling). Las palabras que no se encuentran en el diccionario se muestran en el cuadro de diálogo **Ortografía** que aparece en el lado derecho de la hoja de trabajo.
3. Haga clic en **Ignorar** para cualquiera de los nombres de columna que se identifican como no incluidos en el diccionario.
4. Seleccione **Cambiar Todo** (Change All) para las palabras encontradas mal escritas. (**Cambiar Todo** se encuentra debajo de la flecha desplegable en el extremo derecho del **Cuadro de Sugerencias**.)

¿Cuántas palabras de las columnas que contienen texto se identificaron como incorrectas?

> 1. La palabra Diciembre estaba mal escrita en la celda D19.

Paso 10: Eliminar Formato.

A veces, el conjunto de datos puede tener formato incrustado en la hoja de cálculo. El formato puede ser tan simple como el color o la alineación del texto. También puede haber una condición lógica aplicada a las celdas mediante la opción de formato condicional de Excel que otorga a muchas celdas un formato diferente al resto de la hoja.

En este ejemplo, se eliminará el formato de alineación de la columna **Edad_Grupo** (Age_Group).

1. Seleccione la columna **Edad_Grupo**.
2. Con la columna seleccionada, haga clic en el botón **Borrar** (Clear) en la barra de herramientas **Inicio** (Home) y seleccione **Borrar Formatos** (Clear Formats). Esto eliminará la formación de alineación central de la columna **Edad_Grupo**.

Paso 11: Guarde su Trabajo y Cierre Excel

1. Guarde la hoja de cálculo como Bike Sales_Prepare_Lab 3.4.7_fixed.
2. Cierre Excel.