Laboratorio 2 - Interpretación de visualizaciones con respecto a valores atípicos

Objetivos

En esta práctica de laboratorio, se usarán gráficos y funciones para detectar datos atípicos.

Parte 1: examinar un conjunto de datos en busca de valores atípicos

Antecedentes/Escenario

Un valor atípico es un valor o punto de datos que varía significativamente de otros en el mismo conjunto de datos. Un valor atípico puede resultar de la variabilidad en las mediciones, errores experimentales o errores humanos al ingresar los datos.

Para asegurarse de que cualquier análisis de datos sea correcto, se deben identificar los valores atípicos y luego se debe determinar la mejor manera de tratarlos.

Recursos necesarios

- Dispositivo móvil o PC/computadora portátil con un navegador, Microsoft 365 Excel en línea y acceso a Internet **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 o versión utilizada para lograr los resultados que se muestran en esta práctica de laboratorio.

Instrucciones

Parte 1: Examinar un conjunto de datos en busca de valores atípicos

Paso 1: Abra el conjunto de datos.

1. Descargue el archivo **Bike Sales_Outlier_Lab.xlsx**
2. Cargue el archivo en OneDrive y ábralo en MS 365 Excel en línea.

Paso 2: usar una tabla dinámica para seleccionar los datos para el análisis

1. Haga clic en cualquier celda de la hoja de trabajo de ventas de bicicletas.
2. Inserte una tabla dinámica haciendo clic en **Insertar** (Instert) > **Tabla dinámica** (Pivoy Table). Verifique que Nueva hoja de trabajo esté seleccionada en el cuadro de

diálogo **Crear tabla dinámica** (Create PivotTable) y haga clic en **Aceptar** (OK).

Esto agrega una nueva hoja de trabajo para la tabla dinámica.

3. En el cuadro de diálogo **Campos de tabla dinámica** (PivotTable Fields), marque los campos **Fecha** (Date) y **Cantidad_pedido** (Order_Quantity)

   La tabla dinámica se crea con dos columnas **Fecha** (Date) y **Suma de Cantidad_pedido** (Sum of Order_Quantity.

Paso 3: clasificación de datos para encontrar valores atípicos

Una forma de identificar valores atípicos es simplemente clasificando los datos. Este método funciona con pequeños conjuntos de datos donde los datos se escanean fácilmente.

1. Ordenar la columna **Suma de Cantidad_pedido** de mayor a menor
   1. Seleccione los puntos de datos en la columna **Suma de Cantidad_pedido**. (No seleccione el Total General ni el encabezado de la columna).
   2. Haga clic en **Ordenar y Filtrar** (Sort & Filter) > **Ordenar Descendente** (Sort Descending).

      Esto ordena los puntos de datos **Cantidad_pedido** de mayor a menor.

¿Qué fecha de diciembre tuvo la mayor cantidad de ventas? ¿Cuál fue la cantidad de ventas?

> 19 de diciembre. La cantidad de ventas fue 43.

Revise los datos de la hoja de trabajo Ventas de bicicletas para el 19 de diciembre. ¿Qué entrada contribuye más a la suma de Cantidad_pedido en la tabla dinámica? En otras palabras, ¿Qué número de pedido es el más responsable del valor atípico?

> fila 72 número de pedido 000261765

Paso 4: use un gráfico de dispersión para encontrar valores atípicos

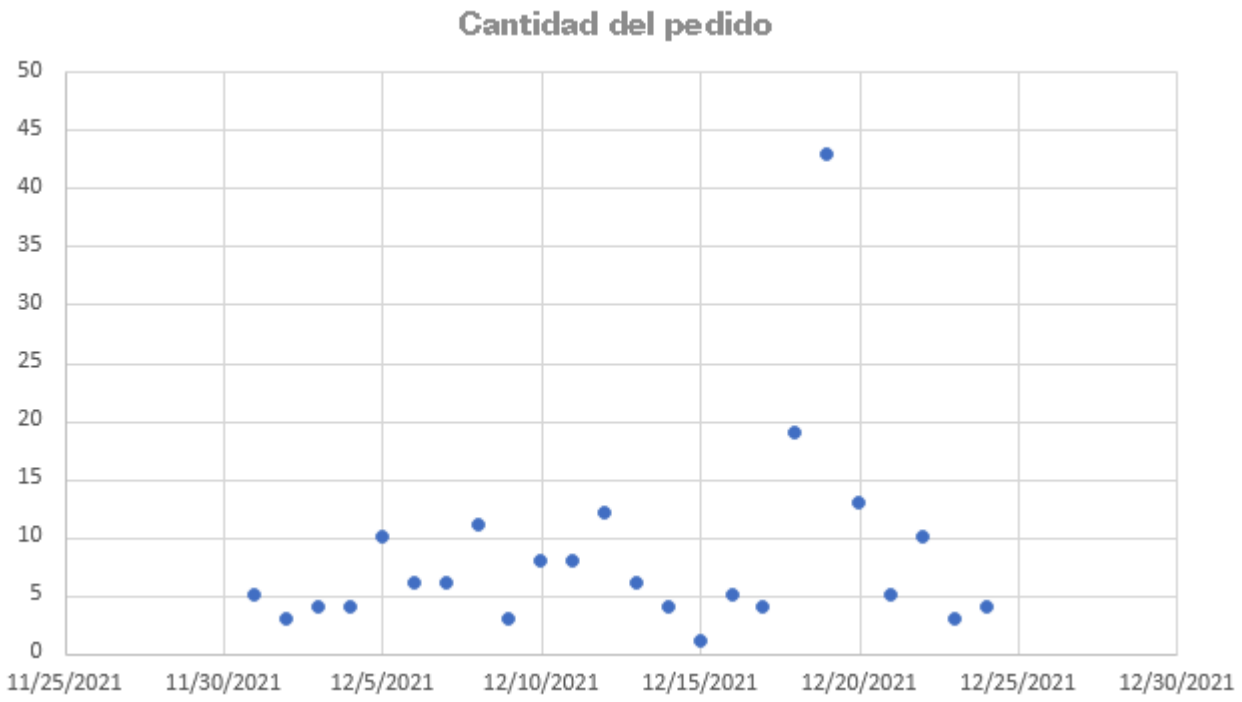
Un gráfico de dispersión puede ayudar a identificar valores atípicos, especialmente en conjuntos de datos más grandes.

1. Regrese a la hoja de trabajo que contiene la tabla dinámica (Hoja1).
2. Copie y pegue los datos de la tabla dinámica en dos columnas en blanco (D y E). Copie la fila del encabezado con los datos, pero no copie la fila del total general.

   Excel no permitirá la creación de un gráfico de dispersión a partir de los datos de una tabla dinámica. Por lo tanto, los datos deben moverse a otras columnas.
3. Insertar gráfico de dispersión.

1. Seleccione todas las celdas de los datos copiados y use Ordenar y filtrar para ordenarlas de manera ascendente.
2. Resalte la columna **Suma de Cantidad_pedido** en los datos copiados.
3. Haga clic en **Insertar** > **Dispersión** y luego seleccione el gráfico de dispersión superior izquierdo en la lista desplegable.

   Tenga en cuenta que lo visual del gráfico de dispersión hace que las ventas del 19 de diciembre se destaquen fácilmente como un valor atípico de los otros puntos de datos de cantidades del pedido, como se muestra a continuación.



4. Elimine el gráfico de dispersión.

Paso 5: Uso de las funciones LARGE (MAX) y SMALL (MIN) para encontrar valores atípicos.

Si hay muchos datos, las funciones LARGE y SMALL se pueden usar para extraer los valores más grandes y más pequeños, lo que puede ayudar a ver si hay valores atípicos.

Para este ejemplo, la columna **Fecha** es la columna D y la columna **Suma de cantidad_pedido** es la columna E. Las columnas de la hoja de trabajo pueden ser diferentes, por lo que debe ajustar las referencias de las celdas de función en consecuencia.

	D	E	F
1			
2			
3	**Date**	**Sum of Order_Quantity**	
4	12/1/2021	5	
5	12/2/2021	3	

1. En una celda vacía, ingrese la función =LARGE(4 :E27,1).

Esta función mira las entradas de la celda E4 a E27 y devuelve el valor más alto.

¿Qué valor se devolvió?

> 43

2. Para obtener los 5 valores más altos, modifique las funciones a **=LARGE(**$4:$ **E27,ROW($1:5))**.

   Esto devuelve los cinco valores más altos. Para devolver más valores, cambie el “5” al final de la función por el número de valores que desea devolver.

¿Qué función devolvería los 6 valores más bajos?

> =K.ESIMO.MENOR(4 :E$27;FILA(1:6))

Una vez que se identifican los valores atípicos, el siguiente desafío es qué hacer con ellos. Los valores atípicos pueden indicar errores en los datos o pueden ser datos válidos que deben investigarse para saber por qué parece ser una anomalía. Hay un par de formas en que un analista de datos puede lidiar con los valores atípicos.

1. Eliminarlos. En un conjunto de datos grande, la eliminación de algunos valores atípicos probablemente no afecte el análisis general. Sin embargo, es importante crear una copia de los datos para poder investigar qué causó los valores atípicos en primer lugar. En este ejemplo, se podría eliminar la fila 72 del conjunto de datos de ventas de bicicletas.
2. Normalícelos (ajuste su valor). El valor de los valores atípicos se cambia para estar ligeramente por encima del valor máximo en el conjunto de datos. Este es un buen método si no sesga los datos. Existen varios métodos estadísticos para normalizar los datos. Investigue los diversos métodos antes de ajustar aleatoriamente los valores de los datos. En el conjunto de datos de ventas de bicicletas de ejemplo, el valor de Order_Quantity del 19 de diciembre de 43 a 20 podría estar justo por encima del valor máximo de 19.

Preguntas de reflexión

Enumere los factores que podrían determinar si los datos atípicos deben considerarse o no en el análisis final de un conjunto de datos.

> 1. El impacto que tienen los datos atípicos en los resultados estadísticos.
> 2. La causa conocida o sospechada del valor atípico (error, rareza, evento real).
> 3. La naturaleza del estudio: exploratorio, confirmatorio o predictivo.
> 4. La cantidad y proporción de datos atípicos respecto al total.
> 5. La relevancia contextual del dato atípico para la toma de decisiones.