# Lab – Interpret Visualizations with Respect to Outliers

## Objectives

In this lab, charts and functions will be used to detect data outliers.

## Part 1: Examine a Dataset for Outliers

### Background / Scenario

An outlier is a value or data point that varies significantly from others in the same dataset. An outlier can result from variability in the measurements, experimental errors, or human error in entering the data.

To make sure that any data analysis is correct, outliers need to be identified and then it needs to be determined how best to treat them.

### Required Resources

- Mobile device or PC/laptop with a browser, Microsoft 365 Excel online, and internet access

**Note**: The precise steps to format and manipulate data in Excel can vary between platforms and versions. The instructions in this lab are based on the free version of Excel available from Office.com and may have to be modified to match the platform or version used to achieve the results shown in this lab.

### Instructions

## Part 1: Examine a Dataset for Outliers

### Step 1: Open the data set.

a. Download the file **Bike Sales_Outlier_Lab.xlsx**

b. Upload the file to your OneDrive and open it in MS 365 Excel online.

### Step 2: Use a Pivot Table to Select Data for Analysis

a. Click any cell in the Bike Sales worksheet.

b. Insert a pivot table by clicking **Insert** > **PivotTable**. Check that New Worksheet is selected in the **Create PivotTable** dialog box and click **OK**.

This adds a new worksheet for the pivot table.

c. In the **PivotTable Fields** Dialog box check the **Date** and **Order_Quantity** fields.

The pivot table is created with two columns **Date** and **Sum of Order_Quantity**.

### Step 3: Sorting Data to Find Outliers

One way to identify outliers is by just sorting the data. This method works with small data sets where the data is easily scanned.

a. Sort the **Sum of Qrder_Quantity** column from high to low

    1. Select the data points in the **Sum of Order_Quantity** column. (Do not select the Grant Total or the column header).

    2. Click **Sort & Filter** > **Sort Descending**.

This sorts the **Order_Quantity** data points from highest to lowest.

Which December date had the largest sales quantity? What was the sales quantity?

¿En qué fecha de diciembre se registró la mayor cantidad de ventas?

## R// 19/12/2021

¿Cuál fue la cantidad de ventas?

## R// 43

Review the data in the **Bike Sales** worksheet for December 19th. Which entry contributes most to the Sum of Order_Quantity in the pivot table? In other words, which order number is most responsible for the outlier?

R// **Fila 72 número de orden 000261765**

**Step 4: Use a Scatter Chart to Find Outliers**

A scatter chart can help to identify outliers, especially in larger datasets.

a. Return to the worksheet containing the pivot table (Sheet1).

b. Copy and paste the data from the pivot table into two blank columns (D and E). Copy the header row with the data, but do not copy the Grand Total row.

Excel will not allow creation of a scatter plot from data in a pivot table. So, the data must be moved to other columns.

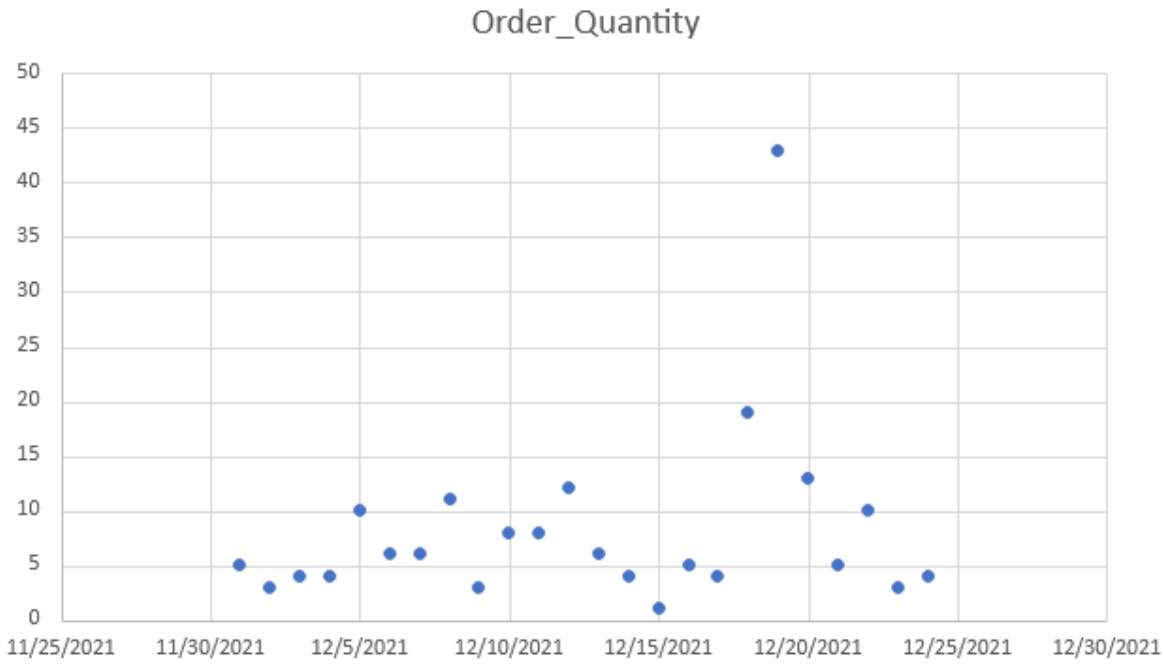
c. Insert scatter plot.

    1. Select the all cells in the copied data and use Sort & Filter to sort it ascending.

    2. Highlight the **Sum of Order_Quantity** column in the copied data.

    3. Click on **Insert** > **Scatter** and then select the top left scatter plot in the dropdown list.

Note that the visual of the scatter chart makes the sales for December 19th easily stand out as an outlier from the other order quantity datapoints as shown below.




4. Delete the scatter plot.

**Step 5: Using the LARGE and SMALL Functions to Find Outliers.**

If there is a lot of data the LARGE and SMALL functions can be used to extract the largest and smallest values which can help to see if there are any outliers.

For this example, the **Date** column is column D and the **Sum of Order_Quantity** column is column E. The columns in your worksheet may be different so adjust your function cells references accordingly.

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

a. In an empty cell enter the function =LARGE($E$4:$E27,1).

This function looks at the entries from cell E4 through E27 and returns the highest value.

What value was returned?

## R// 43

b. To get the highest 5 values, modify the functions to **=LARGE($E$4:$E27, ROW($1:5))**.

This returns the highest five values. To return more values change the “5” at the end of the function to number of values you would like returned.

What function would return the lowest 6 values?

## Respuesta: =PEQUEÑA($E$4:$E27, FILA($1:6))

Once outliers are identified, the next challenge is what to do with them. Outliers may indicate errors in the data, or may be valid data that needs to be investigated as to why it appears to be an anomaly. There are a couple of ways in which a data analyst can deal with outliers.

1. Delete them. In a large dataset deleting a few outliers will likely not impact the overall analysis. However, it is important to create a copy of the data so you can research what was causing the outliers in the first place. In this example, row 72 in the Bike Sales dataset could be deleted.
2. Normalize them (Adjust their value). The value of the outliers is changed to be slightly above the maximum value in the dataset. This is a good method if it will not skew the data. There are a number of statistical methods to normalize data. Research the various methods before randomly adjusting data values. In the example Bike Sales dataset, the December 19$^{th}$ Order_Quantity could be changed from 43 to 20 so it is just above the maximum value of 19.