Not your normal data: outliers

# What are outliers?

Outliers are points that differ significantly form the rest of our observations

Identifying outliers is important because they may:

1. Bias our statistical analysis
2. Signify errors or issues with our data quality
3. Provide insights about interesting or special cases

Outliers need to be examined to asses why they occurred and whether they should be included in the analysis. It doesn’t always mean that these need to be removed, and in fact sometimes outliers require us to give a closer look to determine why it occurred

# How can we identify outliers?

There are quite a few different ways to detect outliers some are very simple vizualizations that only tell you if you have outliers in the data.

Some are very specific calculations to tell you the exact data of outliers.

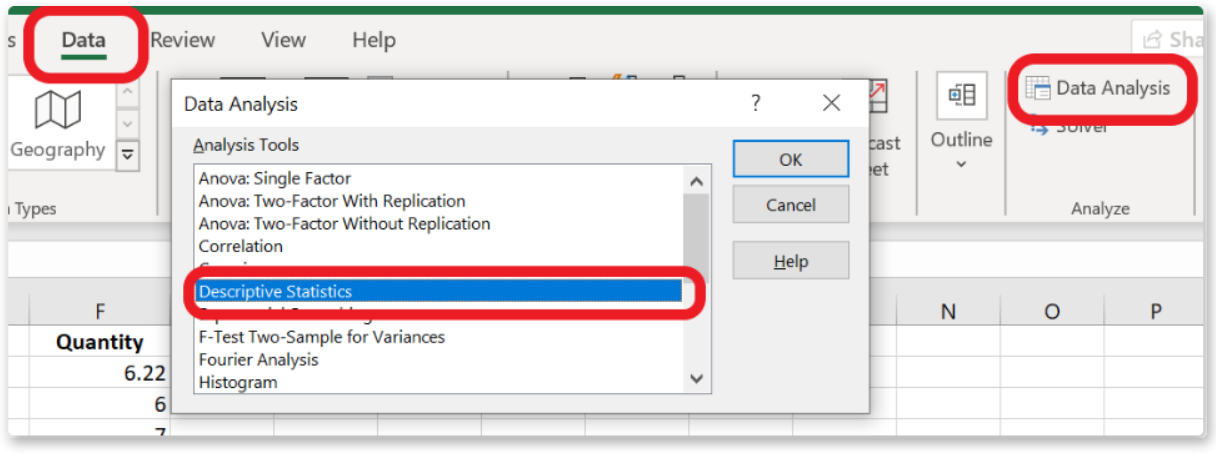
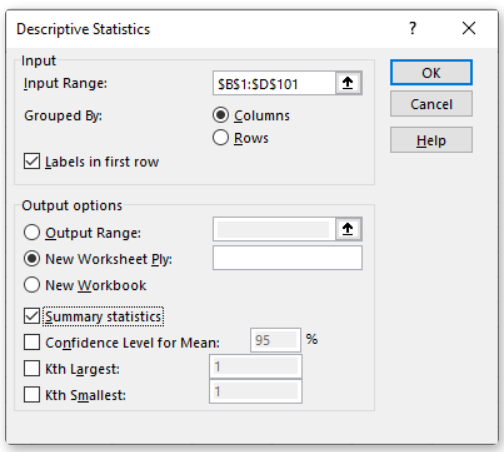
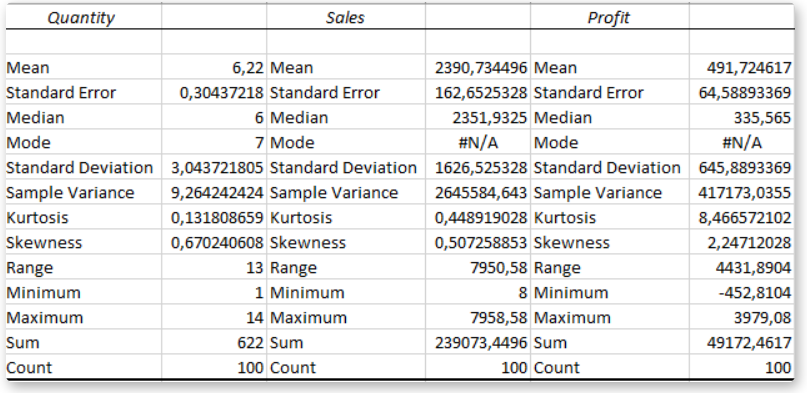
* Mean and standard deviation
* Hitsograms
* Interquartile range and boxplots
* Z-score
* Percentile

# Mean and standard deviation method

In this method we weill uise mean, standard deviation, and specified factors to find out he outliers.

We will use a factor of three here.

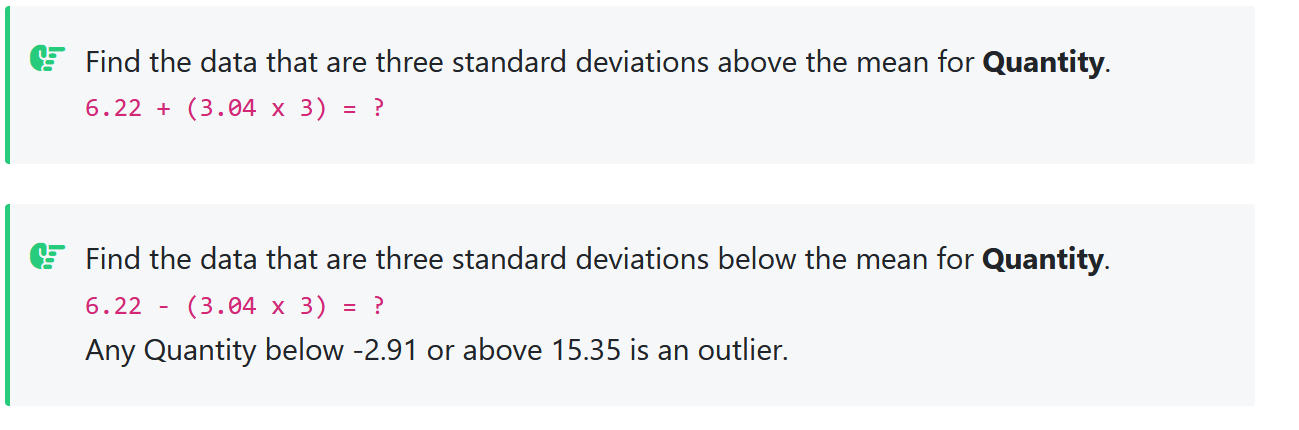
Three standard deviations up from the mean and three standard deviations below the mean will be considered outliers

* We will use the DATA ANALYSIS TOOLPAK to get the standard deviation
* Then, from the DATA menu on the top ribbon, select Data Analysis, followed by Descriptive Statistice from the drop-down list
* Press OK
* 
* Then, in INPUT RANGE: select the data for QUNTITTY, SALES and PROFIT
  + Make sure columns is selected, and tick the box marked LABELS IN FIRST ROW, since our columns include the titles QUANTITY, SALES AND PROFIT
* Then under OUTPUT OPTIONS, make sure that NEW WORKSHEET PLY is selected. We have the noption to save the output to the same worksheet, create a new worksheet or create a new workbook. It is always best practice to split your work into multiple worksheets.
* Tick the box for SUMMARY STATISTICS and press OK
* 
* This gives us the results of the calculations of central tendency and dispersion for Quantity, Sales, and Profit along with some other statistical measures
* 
* For now let’s focus on the measures of central tendancy and dispersion

We are goin to use + OR – 3 deviations from the mean to identify outliers.

The fiormuila is as follows:

* MEAN +/- (STD DEV x3) = ?\*\*

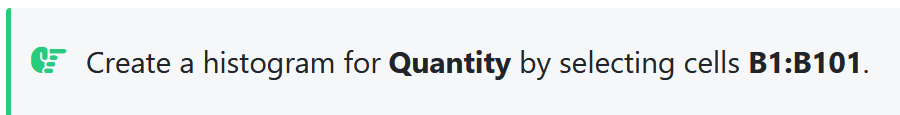


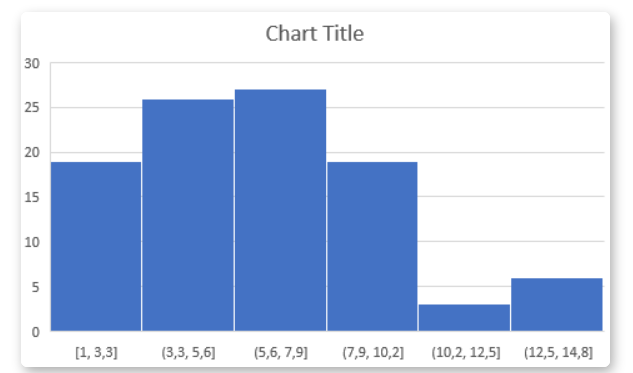
Look at the minimum and maximum of Quantity.

It looks like we don’t have any outliers with a minimum of 1 and a maximum of 14

# Histogram

Looking at he distribution can also tell you if there outliers in the data



* Then from the instert menu, select INSERT STATISTIC CHART, and HISTOGRAM
* 
* The distribution gives the impression that the data is skewed and on the right side there are outliers, but we detect no outliers with the meanand standard deviation formula. It shows that using visualizations to detect outliers is subjective. One can say that there are outliers when in fact there are not. That’s why it is important to use several techniwues to detect them

# interquartile Range

in this method, we need to calculate the first and third quartile to get the interquartile rang (IQR). Then we will consider the first quartile minus 1.5 time IQR as the lower limit and the third quartile plus 1.5 times IQR as the upper limit of the data

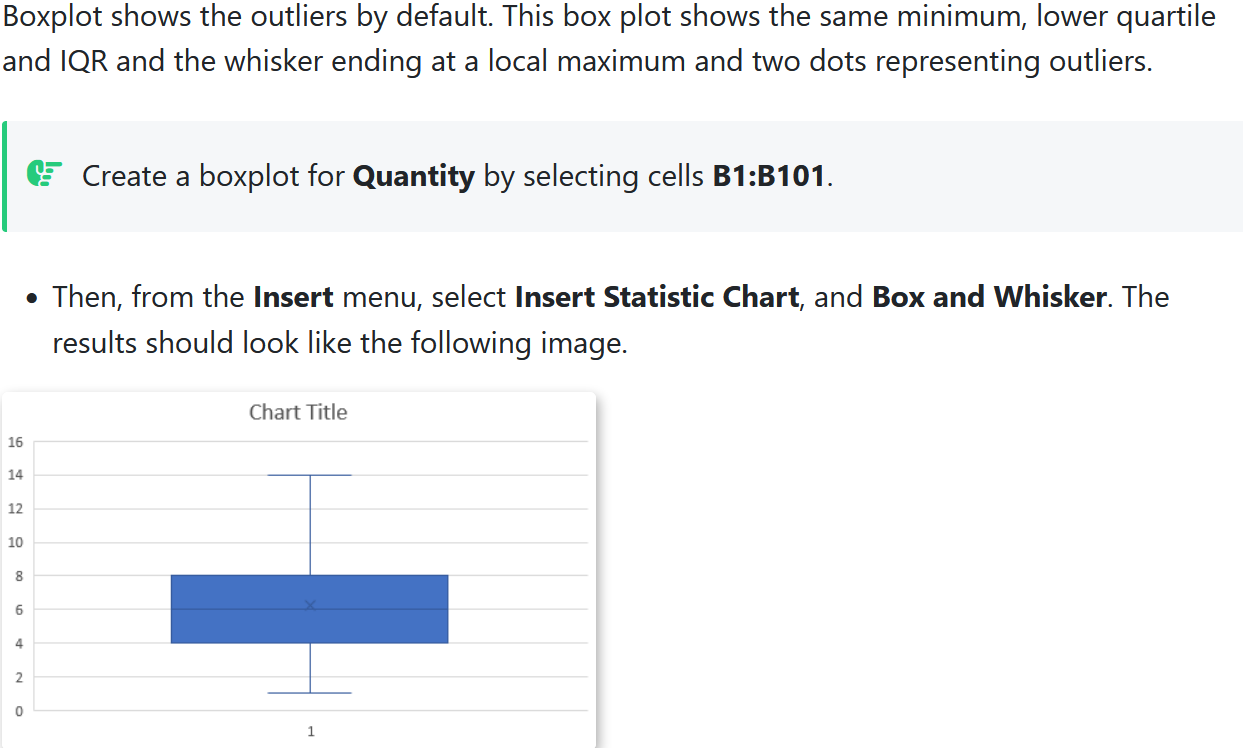
if a value is lower than the lower limit or higher than the upper limit, it will be considered an outlier

to calculate the quartile in Excel for QUANTITY, we can use the QUARTILE FINCTION for both q1 and q3.

* Q1 = QUARTILE(B2:B101,1) -> 4
* Q3 = QUARTILE(B2:B101,3) -> 8
* IQR = Q3-Q1 ->4
* LOWER LIMIT = Q1-(IQR\*1.5) -> -2
* UPPER LIMIT = Q3+(IQR\*1.5) -> 14

Therefore any quantity below -2 or above 14 is an outlier

# Boxplot



# Z-score

Is statistics, a z-score tells us how many SD away a value is from the mean.

We use the following formula to calculate z-score:

Z = (X - µ) / σ , where:

* X is a single raw data value
* µ is the mean of the dataset
* σ is the standard deviation of the dataset
* just fix a z-score threshold and if the z-score is more than that, the data is an outlier

## how to interpret z-scores in excel

a positive z-score tells us that the value associated with this score is above the mean and negative is below the mean and 0 indicates that a particular value is equal to the mean

the first value of Quantity is 2, which has a z-score of -1.4. this means that the value 2 is -1,4 standard deviations below the mean, the next value in our data,9, has a z-score of 0.9.

this means that the value 9 is 0.9 standard deviations above the mean

# percentile calculation

a percentile tells us what percentage of observations fall below a certain value in a data set

you can simply fix a percentile for the upperlimit and lower limit. In this example, we will consider the lower limit as the 10th percentile and the upper limit as the 90th percentile, it will be considered an outlier. To calculate the percentile in excel for quantity, we can use the percentile

* 10th percentile =PERCENTILE(B2:B101,0,1) → 2
* 10th percentile =PERCENTILE(B2:B101,0,9) → 10