# Data Processing: Business Problem

## Outlier

An **Outlier** is an observation or data point significantly different from other observations or data points in a dataset. Various factors can lead to the occurrence of outliers, such as natural variations in the data, measurement errors, or, data entry errors. They can significantly impact data analyses, leading to incorrect conclusions if not identified and dealt with appropriately.

## Missing Data vs Outlier

Missing data and outliers are both issues that can affect the quality of a dataset and the accuracy of the analysis performed on it. However, they are different problems that require distinct approaches to handle them.

**Missing data** refers to the absence of values in a dataset, which can occur for a variety of reasons such as data entry errors or incomplete surveys. Missing data can cause problems such as reduced sample size, biased results, and inaccurate analysis. To handle missing data, different techniques such as imputation, exclusion can be used to fill in the missing values.

On the other hand, **Outliers** are extreme values that deviate significantly from the rest of the data in a dataset. Outliers can be caused by measurement errors, natural variation, or rare events and can skew the results of the analysis by affecting the mean, standard deviation, and other statistical measures.

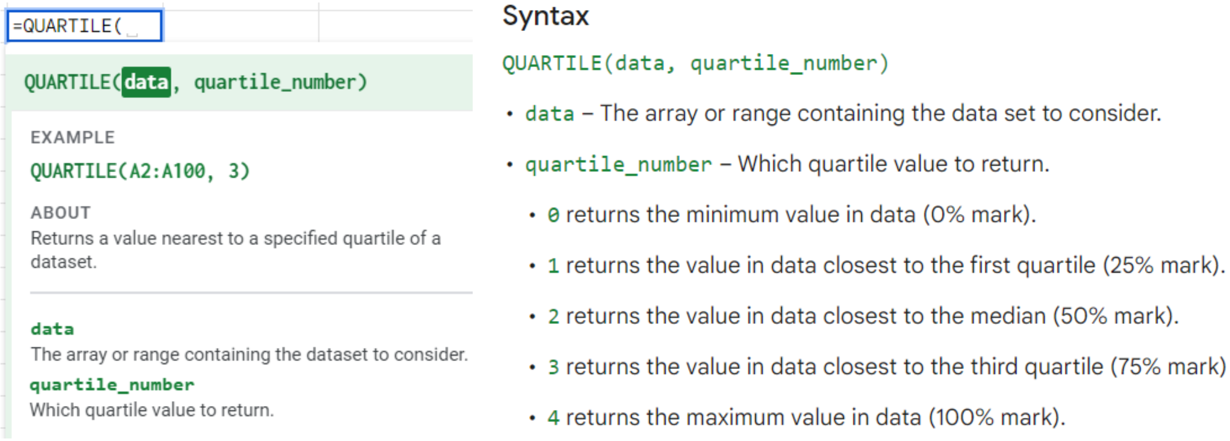
To handle outliers, techniques such as the **Quartile Function** can be used to identify and remove them from the dataset or account for them in the analysis.

# Quartile Function

The **Quartile Function** calculates the quartiles of a dataset, which divides the data into four equal groups. It splits the data range into four sections of equal size. This function can determine the minimum, maximum, first, second, and third quartiles.

**The four quartiles that divide a data set into quartiles are:**

1. The lowest 25% of numbers.
2. The next lowest is 25% of the numbers (up to the median).
3. The second highest 25% of numbers (above the median).
4. The highest 25% of numbers.



```
=QUARTILE(
```

QUARTILE(data, quartile_number)

EXAMPLE

QUARTILE(A2:A100, 3)

ABOUT

Returns a value nearest to a specified quartile of a dataset.

data

The array or range containing the dataset to consider.

quartile_number

Which quartile value to return.

Syntax

QUARTILE(data, quartile_number)

- data – The array or range containing the data set to consider.
- quartile_number – Which quartile value to return.
  - 0 returns the minimum value in data (0% mark).
  - 1 returns the value in data closest to the first quartile (25% mark).
  - 2 returns the value in data closest to the median (50% mark).
  - 3 returns the value in data closest to the third quartile (75% mark)
  - 4 returns the maximum value in data (100% mark).

**Formula:** **QUARTILE (data, quartile_number)**

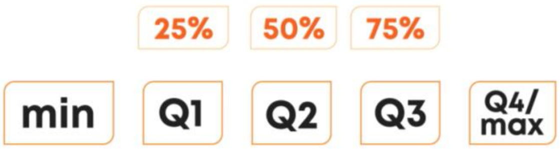
# Steps to find Outlier

**In order to identify outliers, it is necessary to follow these steps:**

**Step 1:** Calculate all the quartile values **(Q1, Q2, Q3)**. For example, to find the quartiles, with a given dataset, {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. The median (Q2) is the middle value, which is 5. Q1 is the median of the lower half of the data, which is (2, 3, 4, 5), Q1 = 3. Q3 is the median of the upper half of the data, which is (6, 7, 8, 9), Q3 = 7.

**Step 2:** Find **Interquartile Range (IQR)** for the required column.
**Formula:** **IQR = Q3 – Q1**

**Step 3:** In the new column apply this **condition:**
**=IF(OR(Cell_valueQ3+1.5IQR),1,0)**

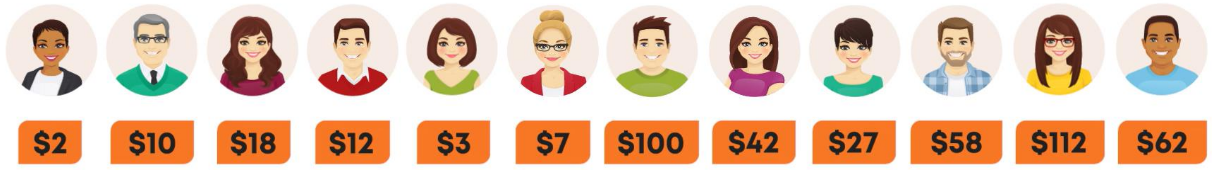


**Step 4:** In the new column, apply the filter or use **=COUNTIF()**

**Step 5:** Calculate the number of outliers present. (Value 1 represents Outlier)

## Example 1

**Given Dataset**: The dataset contains information on 12 customers and the corresponding amount they spent on purchasing fruits.

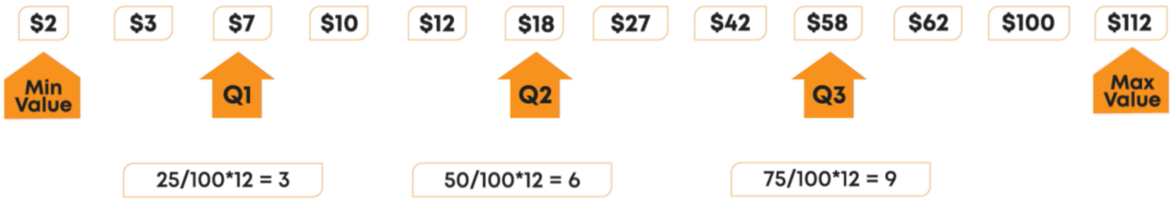


First, sort the given data/column in **increasing order**.

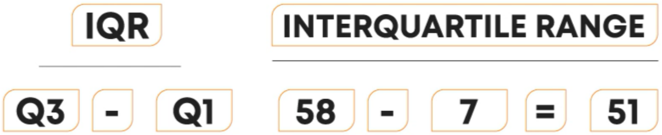
$2 $3 $7 $10 $12 $18 $27 $42 $58 $62 $100 $112

Calculate all the quartile values. Here Q1 = 7, Q2 = 18, Q3 = 58

| min | Q1 | Q2 | Q3 | Q4/ max |
|---|---|---|---|---|
| 2 | 7 | 18 | 58 | 112 |



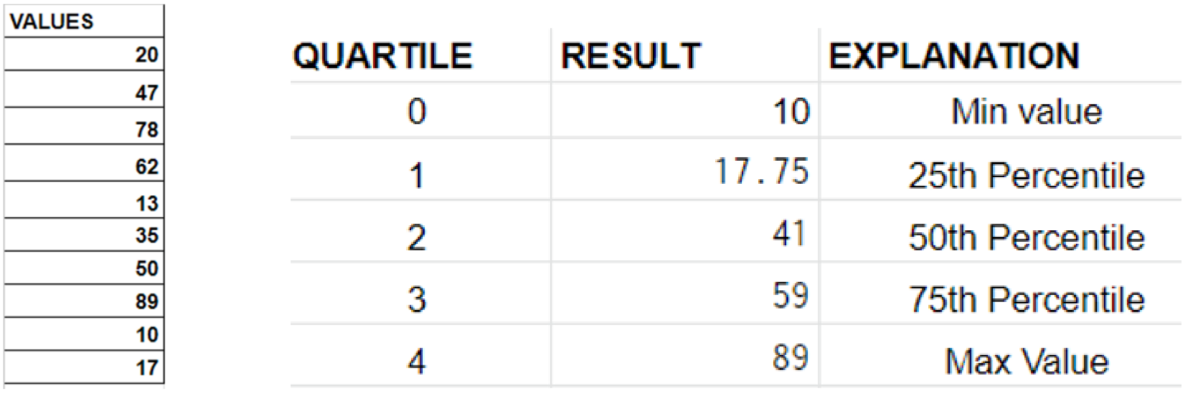
Find the **Interquartile Range (IQR)**. Here IQR = 51



In the new column apply this condition: =IF(OR(Cell_valueQ3+1.5IQR),1,0)
This condition means that,
Value < Q1 – 1.5 * IQR or Value > Q3 + 1.5 * IQR are Outliers.
In the given dataset, data points that fall below -69.5 or above 134.5 will be classified as outliers. You can conclude that there are no outliers present in this given dataset.

## Example 2

Now you will see how to use Google Spreadsheets to find outliers in a given dataset. Here is a sample dataset.



| VALUES |
|---|
| 20 |
| 47 |
| 78 |
| 62 |
| 13 |
| 35 |
| 50 |
| 89 |
| 10 |
| 17 |

| QUARTILE | RESULT | EXPLANATION |
|---|---|---|
| 0 | 10 | Min value |
| 1 | 17.75 | 25th Percentile |
| 2 | 41 | 50th Percentile |
| 3 | 59 | 75th Percentile |
| 4 | 89 | Max Value |

Figure: Sample Dataset and Quartile Values

Here, Qmin = 10, Q1 = 17.75, Q2 = 41, Q3 = 59, Qmax = 89
**IQR=** Q3-Q1 = 59 - 17.75 = 41.25
**1.5 * IQR**= 1.5 * 41.25 = 61.875
**Q1- 1.5*IQR=** 17.75 - 61.875 = -44.125 (Lower Limit)
**Q3 + 1.5*IQR=** 59 + 61.875 = 120.875 (Upper Limit)

In the given dataset, data points that fall below -44.125 or above 120.875 will be classified as outliers. You can conclude that there are no outliers present in this given dataset.

## Box Plot

You can use box plot to find outliers in a dataset. To insert a box plot in Google Spreadsheets:

1. Select the column of data that you want to create a box plot for.
2. Click on the "Insert" menu and select "Chart".
3. In the Chart Editor that appears, select the "Chart types" tab and choose "Box & Whisker chart" from the options.
4. Click "Insert" to add the box plot to your spreadsheet.

Reading a box plot involves interpreting the various components of the plot to understand the distribution of the data.

**Identify the median:** The median is represented by a vertical line inside the box. It indicates the middle value of the data, separating the bottom 50% from the top 50%.

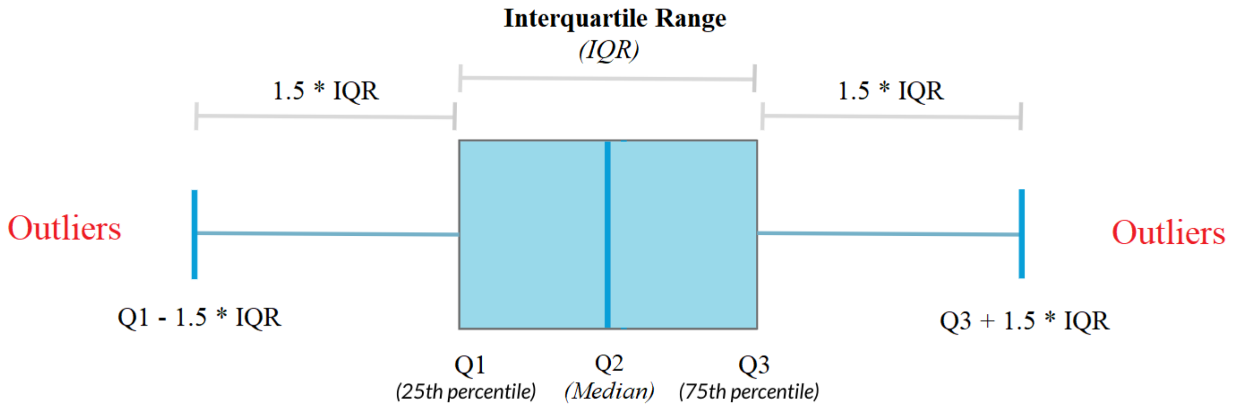


Figure: Identify Outlier in Data

**Determine the interquartile range (IQR):** The IQR is the distance between the first quartile (bottom of the box) and the third quartile (top of the box). It represents the middle 50% of the data.

**Outliers** are represented as individual points outside of the box and lines in a box plot.

## Reference

https://support.google.com/docs/answer/3094041?hl=en-GB