


# How to detect outliers using parametric and non-parametric methods : Part I

 [clevertap.com/blog/how-to-detect-outliers-using-parametric-methods-and-non-parametric-methods](https://clevertap.com/blog/how-to-detect-outliers-using-parametric-methods-and-non-parametric-methods)

An Outlier is an observation or point that is distant from other observations/points. But, how would you quantify the distance of an observation from other observations to qualify it as an outlier. Outliers are also referred to as observations whose probability to occur is low. But, again, what constitutes low??

There are parametric methods and non-parametric methods that are employed to identify outliers. Parametric methods involve assumption of some underlying distribution such as normal distribution whereas there is no such requirement with non-parametric approach. Additionally, you could do a univariate analysis by studying a single variable at a time or multivariate analysis where you would study more than one variable at the same time to identify outliers.

The question arises which approach and which analysis is the right answer??? Unfortunately, there is no single right answer. It depends for what is the end purpose for identifying such outliers. You may want to analyze the variable in isolation or maybe use it among a set of variables to build a predictive model.

Let's try to identify outliers visually.

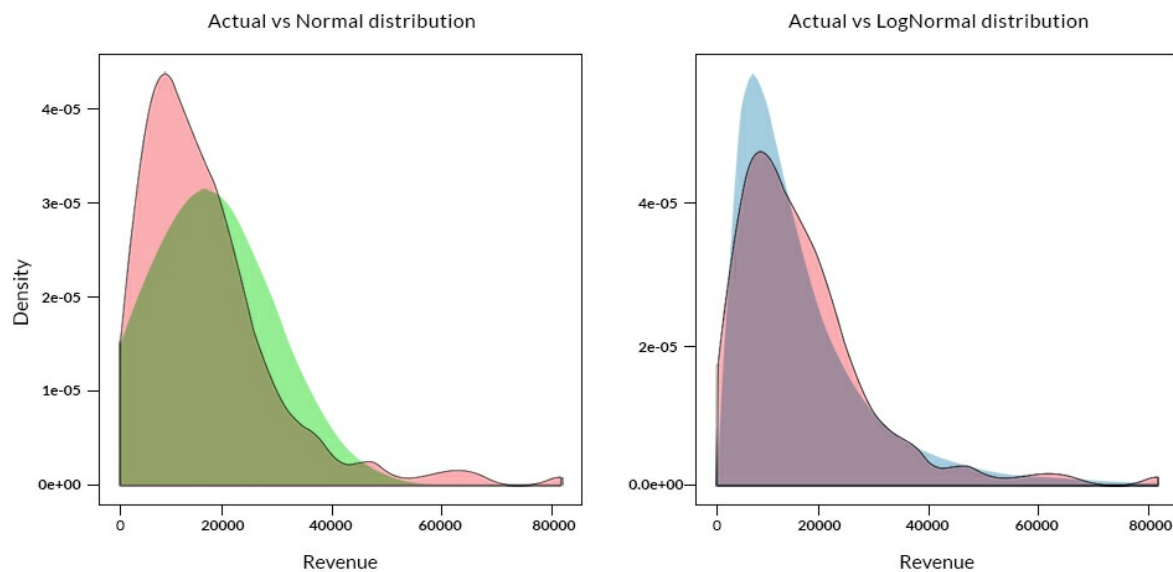
Assume we have the data for Revenue and Operating System for Mobile devices for an app. Below is the subset of the data:

OS	Revenue	Device
Android	8473	Mobile
Android	11790	Mobile
Android	7605	Mobile
iOS	15904	Mobile
iOS	19390	Mobile
iOS	56719	Mobile

How can we identify outliers in the Revenue?

We shall try to detect outliers using parametric as well as non-parametric approach.

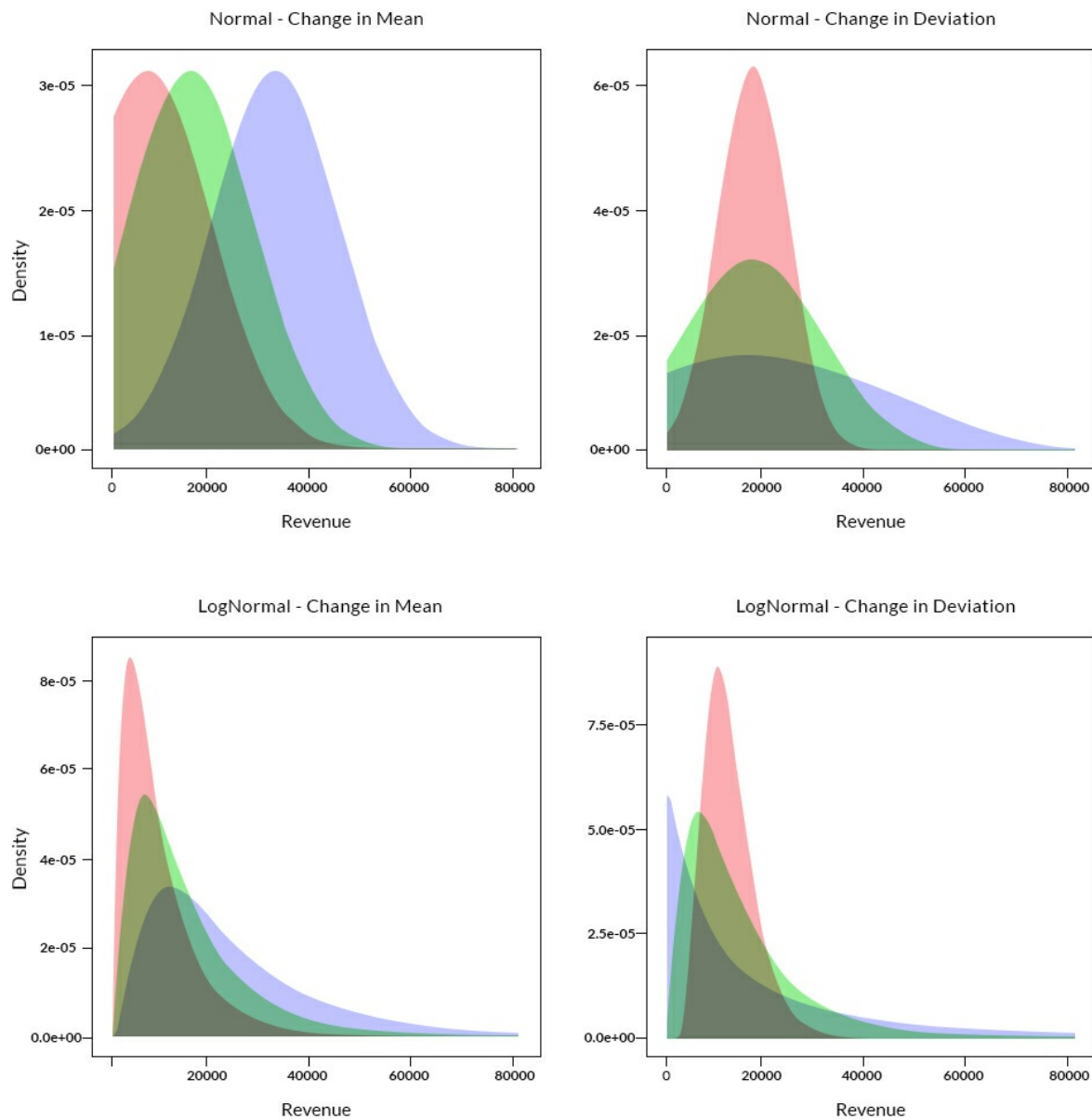
## Parametric Approach



The x-axis, in the above plot, represents the Revenues and the y-axis, probability density of the observed Revenue value. The density curve for the actual data is shaded in 'pink', the normal distribution is shaded in 'green' and log normal distribution is shaded in 'blue'. The probability density for the actual distribution is calculated from the observed data, whereas for both normal and log-normal distribution is computed based on the observed mean and standard deviation of the Revenues.

Outliers could be identified by calculating the probability of the occurrence of an observation or calculating how far the observation is from the mean. For example, observations greater than 3 times the standard deviation from the mean, in case of normal distribution, could be classified as outliers.

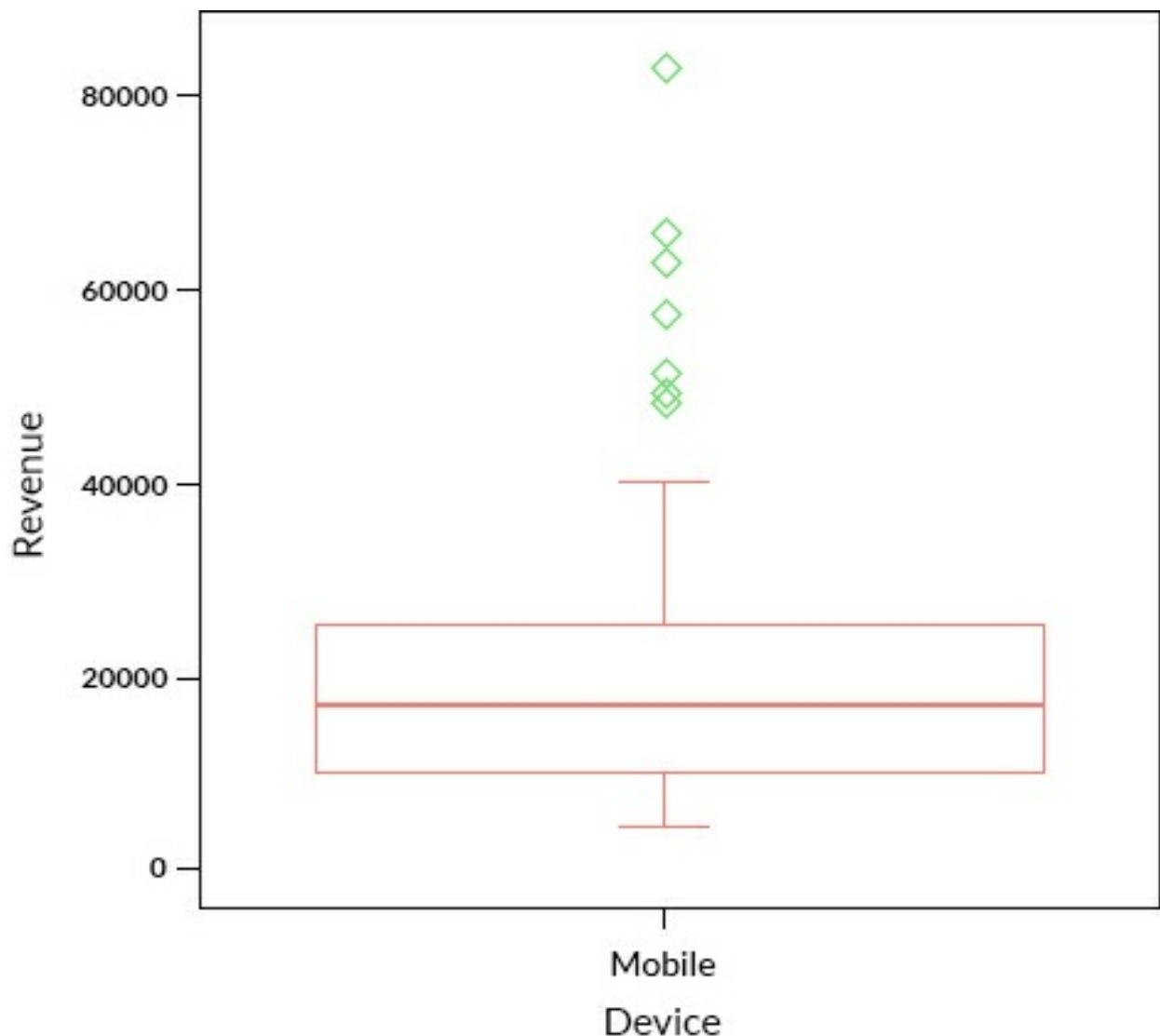
In the above case, if we assume a normal distribution, there could be many outlier candidates especially for observations having revenue beyond 60,000. The log-normal plot does a better job than normal distribution, but it is due to the fact that the underlying actual distribution has characteristics of a log-normal distribution. This could not be a general case since determining the distribution or parameters of the underlying distribution is extremely difficult before hand or apriori. One could infer the parameters of the data by fitting a curve to the data, but a change in the underlying parameters like mean and/or standard deviation due to new incoming data will change the location and shape of the curve as observed in the plots below:



The above plots show the shift in location or the spread of the density curve based on an assumed change in mean or standard deviation of the underlying distribution. It is evident that a shift in the parameters of a distribution is likely to influence the identification of outliers.

### Non-Parametric Approach

Let's look at a simple non-parametric approach like a box plot to identify the outliers.



In the box plot shown above, we can identify 7 observations, which could be classified as potential outliers, marked in green. These observations are beyond the whiskers. ([Read this article to know more about box plots](#)).

In the data, we have also been provided information on the OS. Would we identify the same outliers, if we plot the Revenue based on OS??

#### Non Parametric approach to detect outlier with box plots (bivariate approach)


In the above box plot, we are doing a bivariate analysis, taking 2 variables at a time which is a special case of multivariate analysis. It seems that there are 3 outlier candidates for iOS whereas there are none for Android. This was due to the difference in distribution of Revenues for Android and iOS users. So, just analyzing Revenue variable on its own i.e univariate analysis, we were able to identify 7 outlier candidates which dropped to 3 candidates when a bivariate analysis was performed.

### Closing Thoughts

Both Parametric as well as Non-Parametric approach could be used to identify outliers based on the characteristics of the underlying distribution. If the mean accurately represents the center of the distribution and the data set is large enough, parametric approach could be used whereas if the median represents the center of the distribution, non-parametric approach to identify outliers is suitable.

In Part II, we shall explore outliers by conducting a multivariate analysis with clustering, a popular data mining technique.

# How to detect outliers using parametric and non-parametric methods : Part II

 [clevertap.com/blog/how-to-detect-outliers-using-parametric-and-non-parametric-methods-part-ii](https://clevertap.com/blog/how-to-detect-outliers-using-parametric-and-non-parametric-methods-part-ii)

In the [previous article](#), we discussed what an outlier is and ways to detect such outliers with parametric and non-parametric methods by conducting a univariate and bivariate analysis. Let's now look at Clustering, a non-parametric method and a popular data mining technique to detect such outliers when we are dealing with many variables or in a multivariate scenario.

**Clustering** groups set of objects in such a way that objects or observations in the same group or cluster are more similar to each other, than those in other groups or clusters. The important term here is 'similar'. Somehow, we need to quantify 'similar' in the context of clustering. This could be quantified with a distance measure. Different Clustering techniques employ different distance measures to arrive at the goal of forming clusters; we shall concentrate on K-means clustering, a popular and most widely used clustering technique.

**K-means** clustering is *implicitly based* on pairwise *Euclidean* distances. Euclidean distance is the "ordinary" (i.e. straight-line) distance between two points. There are other distance metrics like Manhattan, Minkowski, Chebychev distance apart from Euclidean distance that can be used but K-means may stop converging with other distance functions.

Given a set of observations  $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ , where each observation is a  $d$ -dimensional real vector,  $k$ -means clustering aims to partition the  $n$  observations into  $k(\leq n)$  sets  $(\mathbf{S} = \{S_1, S_2, \dots, S_k\})$  so as to minimize the within-cluster sum of squares (wss) (sum of distances of each point in the cluster to the  $K^{\text{th}}$  center). In other words, its objective is to find:

where  $\mu_i$  is the mean of points in  $S_i$ .

The above formula will become clear as we discuss the basic steps in K-means clustering and the evaluation metric for identifying the right number of clusters with it.

$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

## Basic Steps in K-means Clustering

### Step 1: Deciding the Number of Clusters

Select the number of clusters or segments. There is no magic formula to calculate this before hand and it has to be an iterative process. We will discuss ways and means to evaluate it with an example.

### Step 2: Assigning Centroids

Randomly select the number of observations or points, which are equal to the number of

clusters. So if we have decided on having 5 segments, randomly select 5 observations from your dataset and assign them as Centroids.

### Step 3: Distance from Centroids

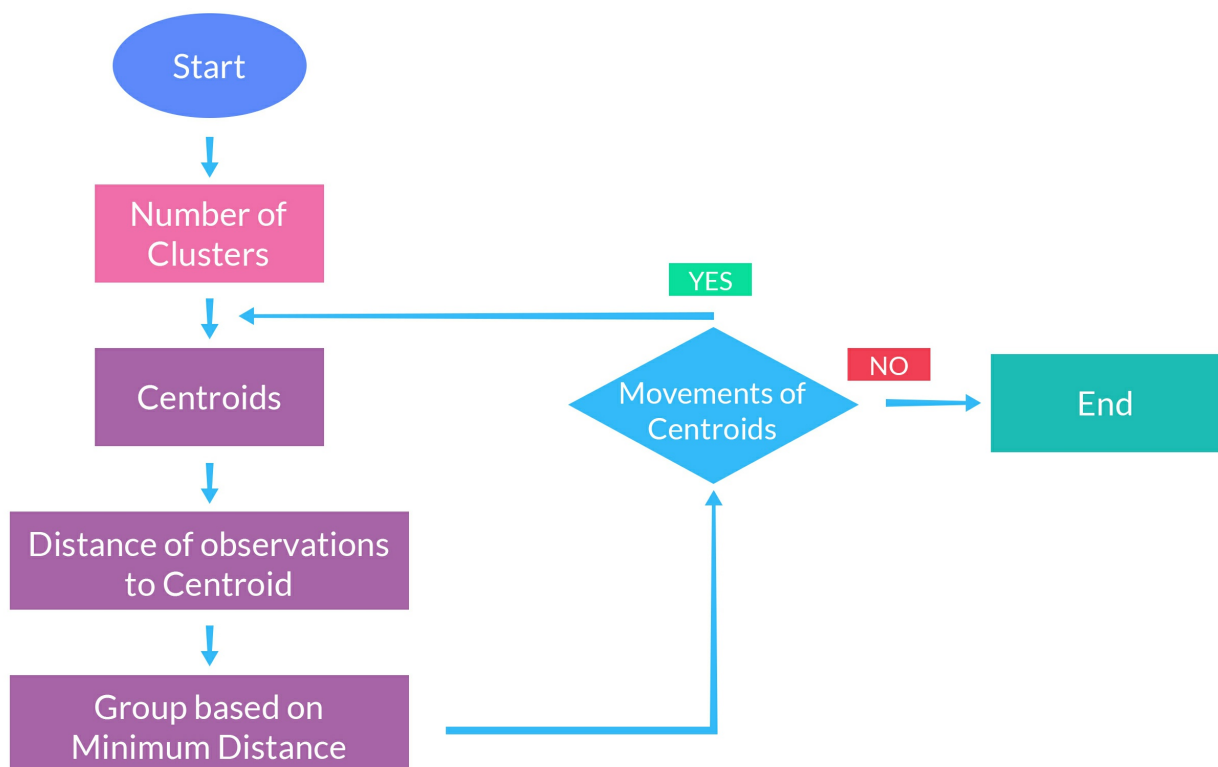
Calculate the distance of each of the observation from the centroids and assign the points to the closest centroid. For example: In a dataset of 100 observations, you would calculate the distance of each of those observations from the centroids. For a particular observation and 5 centroids, you will end up with 5 distances. The observation gets assigned/grouped to the cluster containing the centroid from which it has recorded the lowest distance.

### Step 4: Recalculate Centroids

The center point or the centroid of the clusters is recalculated. The centroid is like the mean of the observations/points and is calculated in the same way as you calculate the average or the mean of points. So, if the points in the cluster change due to assignment of the observations to different clusters in Step 3, the mean or the centroid of that cluster could change.

### Step 5: Repeat

Repeat Step 3 and Step 4 until the Centroids don't move.



## Pre-Processing

One important point to bear in mind before running the K-means algorithm is to standardize the variables. For example, if your dataset consists of variables like age, height, weight, revenue, etc which are essentially incomparable, it will be a good idea to bring them on the same scale. Even if variables are of the same units but with quite different variances it is better to standardize before running the K-means algorithm. K-means clustering is

“isotropic” in all directions of space and therefore, tends to produce more or less round (rather than elongated) clusters. In this situation, leaving variances unequal is equivalent to putting more weight on variables with smaller variance, so clusters will tend to be separated along variables with greater variance. Standardization will ensure that K-means gives equal weightage to all the variables. A quick way to standardize a variable is to take the difference of each observation from the mean of the variable and divide such difference by the standard deviation of the variable.

## Detect Outliers with K-means

Now let's use K-means to identify outliers. This is possible assuming you have the right number of clusters and one of the clusters, due to its characteristics is totally or substantially different from other clusters. This will become clear with an example given below:

User	OS	Avg. Visits	Avg. Transactions	Avg. Revenue
A	Android	10	1	48000
B	iOS	15	4.11	67000
C	Android	4.2	2	40566
D	Android	14	3.25	65113
E	Android	6.1	1.1	40112
F	iOS	7.5	2.33	53800

The above table is a subset of the user data of an app. We have 5 columns/variables with 3 numerical variables and 2 categorical variables. We will ignore Column 1 for our analysis.

### Use of Categorical Variables in K-means

We could convert the categorical variable ('OS') to numerical by encoding Android as '0' and iOS as '1' or vice-versa. But, conceptually it is not advisable to do so and run a K-means algorithm on it due to the manner in which K-means calculates the distance in the Euclidean space. Can you quantify distance between a Android and iOS as 1 or 2 or any other number just as we can easily quantify the distance between 2 numerical points ?

For example: if the Euclidean distance between numeric points A and B is 30 and A and C is 8, we know A is closer to C than B. Categorical values are not numbers but are enumerations such as 'banana', 'apple' and 'oranges'. Euclidean distance cannot be used to compute distances between the above fruits. We cannot say apple is closer to orange or banana because Euclidean distance is not meant to handle such information. For this reason, we won't be including the categorical variable in the K-means algorithm. For categorical variables, a variation of K-means known as K-modes, introduced in [this paper](#) by Zhexue Huang, could be suitable.

### Estimating Number of Clusters



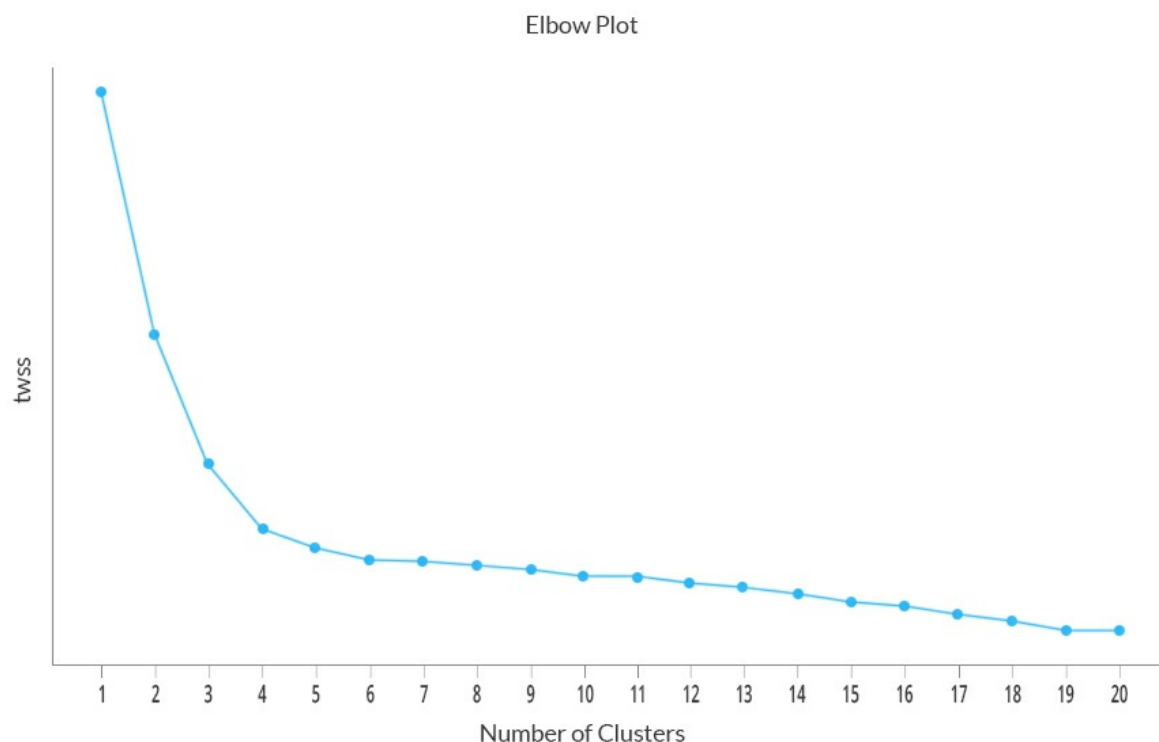
Getting the right number of clusters is an iterative process. To get the right number of clusters, we shall try running the K-means with 2 to 20 clusters for the above example and look at some common evaluation metric to arrive at that number. As an analyst, running the iterative process over higher or lower number of clusters is your prerogative. But, before running the K-means algorithm, we will standardize the numerical variables to bring them on the same scale.

To evaluate the ideal cluster size, there are various metrics like within sum of squares, between sum of squares, percentage of variance explained by the clusters, gap statistic, etc. For the above example, we shall track the within sum of squares.

Within sum of squares (wss):

This metric essentially gives us the homogeneity within a group. The metric calculates the sum of squares of the distance of each point in a particular cluster from its respective centroid. We shall refer to it as 'wss'. You will calculate 'wss' for each of the cluster. The sum total of wss for all the clusters is the Total wss (twss).

The metrics mentioned above are relative and not absolute metrics i.e. their values have to be compared with the number of clusters chosen. The metric for a particular cluster number, say 2, in the above case, can't be viewed in isolation. It has to be compared with all the clusters ranging from 3 to 20. If we have to choose between 2 and 20 as the ideal number of clusters, we will tend to go with the cluster number that has the lowest twss in the above case.



The above graph plots the total within sum of squares (twss) for number of clusters between 1 and 20. From the graph, it seems that the rate of change for twss decreases substantially after 4 clusters. Theoretically, you would tend to choose the lowest twss, but practically it might not be the best option. As per the graph, it seems that the right number of clusters is 19 but analyzing such a large number of cluster tends to be tedious and may

not achieve the required goal as the distance between many clusters may not be significant and could be close to each other. In this example, we see that there is a steep decline in twss until 4 clusters, after which the rate of change decreases rapidly. Hence, we shall go with 4 as the number of clusters. You might even consider analyzing with different cluster sizes and compare the cluster characteristics between all such chosen clusters before arriving at the best cluster size.

## Visualizing Clusters

### Cluster Analysis

The different colors for the points indicate their membership for different clusters. There is one cluster i.e. Cluster 4 that is very different or far from the other clusters, the points for which are in colored in 'blue'.

### Cluster Characteristics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Size	6823	9370	12692	1115	30000
%	22.7%	31.2%	42.3%	3.7%	100.0%

The above table gives the size or the number of observations in each cluster. Cluster 4 has 3.7% of the observations and is the smallest of the lot.

Let's look at some more characteristics of the clusters formed.

Cluster 1					Cluster 2				
	CM	PM	PSd	z score		CM	PM	PSd	z score
Average Visits	9	7.9	3.1	0.35	Average Visits	7	7.9	3.1	-0.29
Average Transactions	1.5	0.91	0.5	1.18	Average Transactions	0.6	0.91	0.5	-0.62
Average Revenue	52149	47788	2439	1.79	Average Revenue	42143	47788	2439	-2.31

Cluster 3					Cluster 4				
	CM	PM	PSd	z score		CM	PM	PSd	z score
Average Visits	7.4	7.9	3.1	-0.16	Average Visits	14.4	7.9	3.1	2.10
Average Transactions	0.8	0.91	0.5	-0.22	Average Transactions	1.16	0.91	0.5	0.50
Average Revenue	48010	47788	2439	0.09	Average Revenue	66013	47788	2439	7.47

CM – Cluster Mean ; PM – Population Mean ; PSd – Population Standard Deviation

The above table shows the mean of the variables observed in the cluster and compares it to the overall mean observed for those variables and uses a Z-score to quantify the distance between the cluster and the overall mean.

$$z = \frac{(x - \mu)}{\sigma}$$

where:  $\mu$  is the mean of the population and  $\sigma$  is the standard deviation of the population.

For each numerical variable, the Z-score, basically, normalizes the distance of the cluster mean from the overall mean by dividing such distance with its standard deviation. For the same mean difference, the variable that has a lower standard deviation will have a higher Z-score between the 2 variables. The sign of the Z-score indicates whether the group mean has been above or below the overall mean.

Cluster 4 seems to stand out in the above table considering factors such as lower number of observations, higher number of visits, transactions, average transaction value, and high Z-scores.

Let's also take a quick look at the cluster characteristics in terms of OS, the categorical variable.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Overall
Android	65%	76%	71%	39%	70%
iOS	35%	24%	29%	61%	30%

The proportion of iOS users in Cluster 4 is significantly different from all the other clusters. Thus, it could be concluded based on the above evidences that Cluster 4 is an outlier and observations in it could be considered as outlier candidates. As an analyst, one should delve deeper into Cluster 4 to understand the key factors driving the behavior of the users in this cluster since its users have not only interacted more with the app but also spent more compared to users in other clusters.

In conclusion, we have been able to identify outliers in a multivariate scenario with the help of clustering, a non-parametric approach. There are a number of clustering algorithms apart from K-means that one can experiment to identify such outliers, based on the characteristics of the dataset, but is beyond the scope of this article.