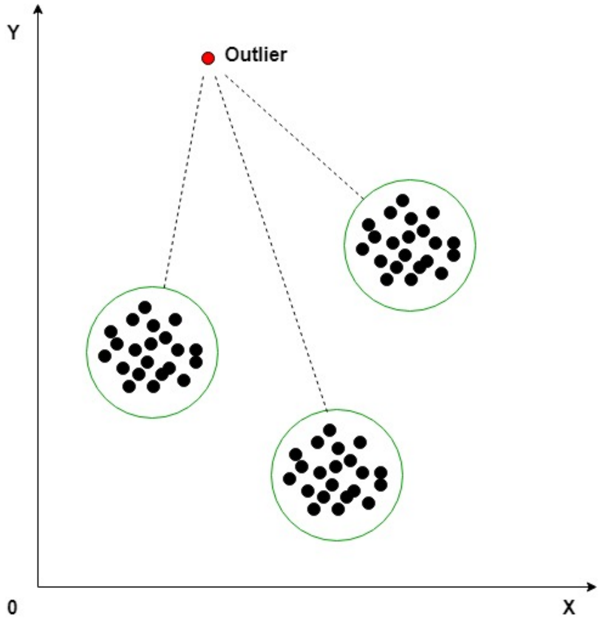
A Brief Overview of Outlier Detection Techniques

"*Observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism" — Hawkins(1980)*

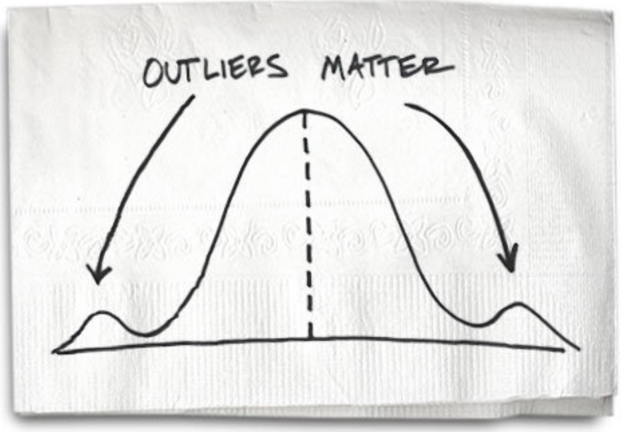
An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Why outlier analysis?

Most data mining methods discard outliers noise or exceptions, however, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring one and hence, the outlier analysis becomes important in such case.



Another Way you can explain Outlier:



- Outliers are extreme values that deviate from other observations on data ,
- They may indicate a variability in a
  - measurement,
  - experimental errors or
  - a novelty
    - In the process of producing, collecting, processing and analyzing data, outliers can come from many sources and hide in many dimensions. Those that are not a product of an error are called **noveltie**s..
- In other words, an outlier is an observation that diverges from an overall pattern on a sample.

Most common causes of outliers on a data set:

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or data set unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

Outlier detection is important in many applications, such as:

- Intrusions in communication networks
- Fraud in financial data
- Fake news and misinformation
- Healthcare analysis
- Industry damage detection
- Security and surveillance
- etc

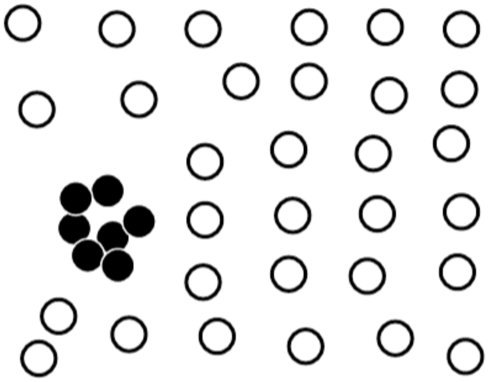
Categorization of outliers

1. Univariate and Multivariate

- Univariate outliers can be found when looking at a distribution of values in a single feature space.
- Multivariate outliers can be found in a n-dimensional space (of n-features). Looking at distributions in n-dimensional spaces can be very difficult for the human brain, that is why we need to train a model to do it for us.

2. Depending on the environment:

- Global outlier (Point outlier)— Object significantly deviates from the rest of the data set
- **Contextual outliers**: Object deviates significantly based on a selected context.
  - noise in data, such as punctuation symbols when realizing text analysis
  - background noise signal when doing speech recognition
  - For example, 28   C is an outlier for a Moscow winter, but not an outlier in another context, 28   C is not an outlier for a Moscow summer
- **Collective outliers**: Usually, a data set may contain different types of outliers and at the same time may belong to more than one type of outlier.
  - A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers. For example, a large set of transactions of the same stock among a small party in a short period can be considered as an evidence of market manipulation



- Which and how many features am I taking into account to detect outliers ? (**univariate** / **multivariate**)
- Can I assume a distribution(s) of values for my selected features? (**parametric** / **non-parametric**)

Will lead to different types of Detection methods

Outlier Detection Methods

- The outlier detection methods differ according to
    - **Statistical study regarding normal objects versus outliers**
    - **Proximity Based**
        - **Density based**
        - **Distance based**
        - sample of data for analysis is given with domain expert–provided labels that can be used to build an outlier detection model.
        - If expert-labeled examples of normal and/or outlier objects can be obtained, they can be used to build outlier detection models.
        - The methods used can be divided into supervised methods, semi-supervised methods, and unsupervised methods.

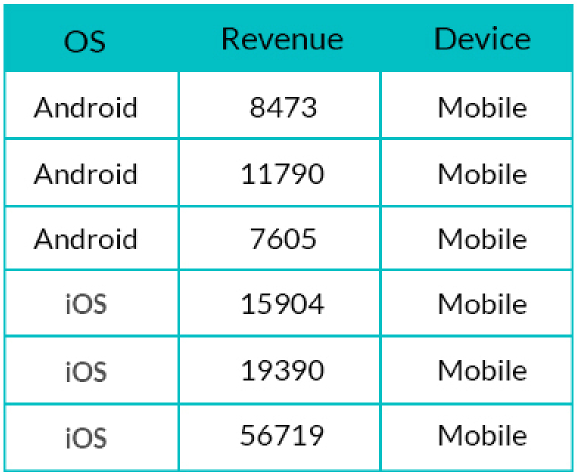
Statistical Methods

Parametric and Non parametric

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.

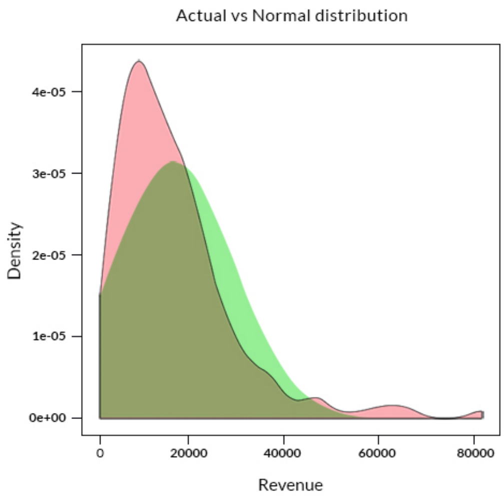
Assume we have the data for Revenue and Operating System for Mobile devices for an app. Below is the subset of the data:



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

Problem : How can we identify outliers in the Revenue?

Parametric Approach



In above Graph :

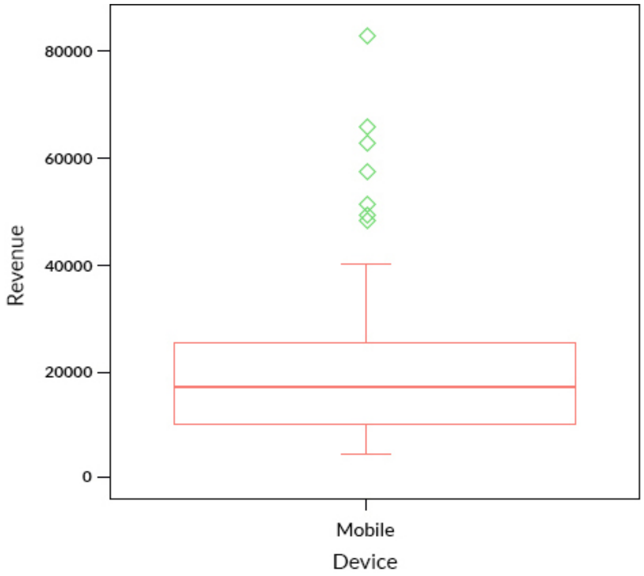
- The x-axis, in the above plot, represents the Revenues and the y-axis, probability density of the observed Revenue value.
- The density curve for the actual data is shaded in ‘pink’,
- the normal distribution is shaded in ‘green’ and
- log normal distribution is shaded in ‘blue’.

- Note :

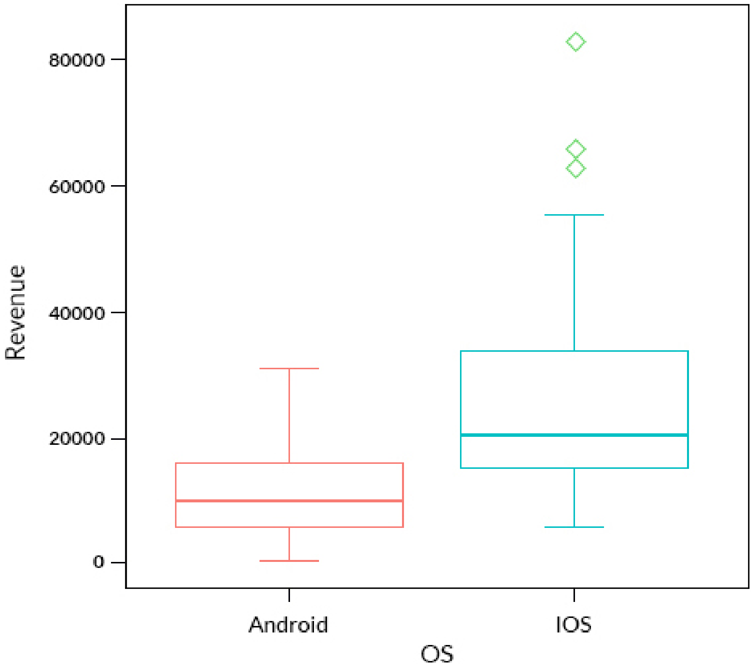
  - The probability density for the actual distribution is calculated from the observed data,
  - whereas for normal distribution is computed based on the observed mean and standard deviation of the Revenues.
- Outliers could be identified by calculating the probability of the occurrence of an observation 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.

Non-Parametric Approach

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



In the box plot plot shown above, we can identify 7 observations, which could be classified as potential outliers, marked in green. These observations are beyond the whiskers.
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??



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.

Some of the most popular statistical methods for outlier detection are:

- **Z-Score or Extreme Value Analysis (parametric)**
- Probabilistic and Statistical Modeling (parametric)
- Linear Regression Models (PCA, LMS)
- **Proximity Based Models (non-parametric)**
- Information Theory Models
- High Dimensional Outlier Detection Methods (high dimensional sparse data)

Z-Score

- The z-score or standard score of an observation is a metric that indicates how many standard deviations a data point is from the sample's mean, assuming a gaussian distribution (If not in normal distribution, proper transformation is to be applied).

- This makes z-score a parametric method.

- z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the **population mean** a raw score is.

- A z-score can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the mean μ (Population Mean) and standard deviation σ.

- After making the appropriate transformations to the selected feature space of the dataset, the z-score of any data point can be calculated with the following expression:

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

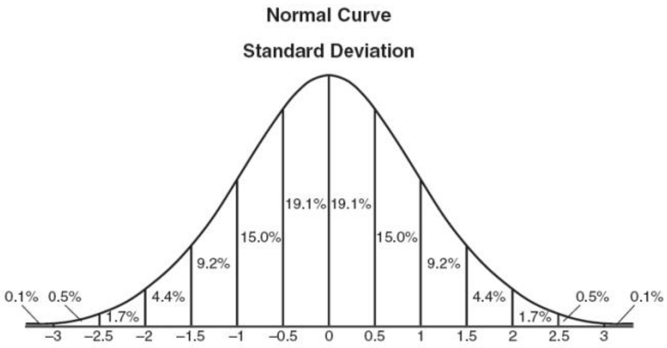
- E.g. For the test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be:

$z = (x – \mu) / \sigma$

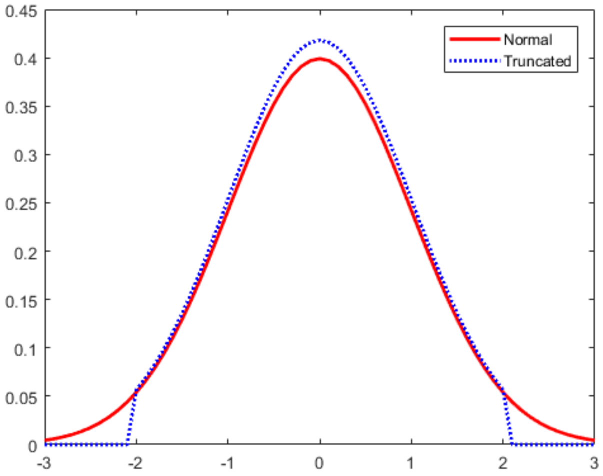
$= (190 – 150) / 25 = 1.6.$

- The z score tells you how many standard deviations from the mean your score is. In this example, your score is 1.6 standard deviations *above* the mean.

- When computing the z-score for each sample on the data set a threshold must be specified. Some good 'thumb-rule' thresholds can be: 2.5, 3, 3.5 or more standard deviations.



By 'tagging' or removing the data points that lay beyond a given threshold we are classifying data into outliers and not outliers



Z-score is a simple, yet powerful method to get rid of outliers in data if you are dealing with parametric distributions in a low dimensional feature space.

Population mean (weighted mean):

Population Mean Definition

- The population mean is an average of a group (population) characteristic. The group could be a person, item, or thing,
- In statistics, it's actually rare that you can calculate the population mean. That's because asking an entire population about something is usually cost prohibitive or too time consuming.
- For example, one veterinary practice probably keeps weight records of all the pets that come in the door, enabling you to calculate the average weight of a dog for that practice (i.e. the population mean for that practice).
- But if you were working for a pet food company who wanted to know the average weight of a dog, you wouldn't be able to track down all the 70 to 80 million dogs in the US and weigh them.
- You would have to take a sample (a small portion of the population of dogs) and weigh them. You can then use this figure to *approximate* the population mean.
- The population mean symbol is μ.
- formula to find the population mean is:
  $\mu = (\Sigma * X)/ N$
  where:
  Σ means "the sum of."
  X = all the individual items in the group.
  N = the number of items in the group.
- **Sample question:** All 57 residents in a nursing home were surveyed to see how many times a day they eat meals.
  1 meal (2 people)
  2 meals (7 people)
  3 meals (28 people)
  4 meals (12 people)
  5 meals (8 people)
  What is the population mean for the number of meals eaten per day?

Solution:
Step 1: Sum up all of your X values. This is the ΣX portion of the population mean formula.
1 + 1 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 5 + 5 + 5 + 5 + 5 + 5 + 5 + 5 = 188.
Note: You could also sum this with the following formula:
(1*2)+(2*7)+(3*28)+(4*12)+(5*8)=188.

- Step 2: Divide your answer to Step 1 with the number of items in your data set. There are 57 people, so:
  188 / 57 = 3.29824561404
  That’s an average of 3.3 meals per person, per day.
  The population mean is 3.3.

- **Population Mean vs. Sample Mean**

- Figuring out the population mean should feel familiar. You’re just taking an average, using the same formula you probably learned in basic math (just with different notation). However, care must be taken to ensure that you are calculating the mean for a population (the whole group) and not a sample (part of the group). The symbols for the two are different:
  Population mean symbol = μ
  Sample mean symbol = $\bar{x}$

Nonparametric problems (Dbscan)

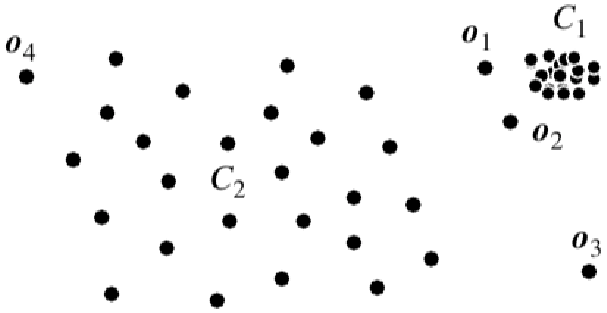
In nonparametric methods for outlier detection, the model of “normal data” is learned from the input data, rather than assuming one a priori. Nonparametric methods often make fewer assumptions about the data and thus can be applicable in more scenarios. As an example, we can use histograms.

Distance Based :

- ❖ Distance-based outlier detection method consults the neighbourhood of an object, which is defined by a given radius. An object is then considered an outlier if its neighborhood does not have enough other points.
- ❖ A distance the threshold that can be defined as a reasonable neighbourhood of the object.

Density Based :

- ❖ Density-based outlier detection method investigates the density of an object and that of its neighbors. Here, an object is identified as an outlier if its density is relatively much lower than that of its neighbors.



- ❖ Consider the example above, distance-based methods are able to detect O_3, but as for O_1 and O_2 it is not as evident.
- ❖ The idea of density-based is that we need to compare the density around an object with the density around its local neighbors. The basic assumption of density-based outlier detection methods is that the density around a nonoutlier object is similar to the density around its neighbors, while the density around an outlier object is significantly different from the density around its neighbors.

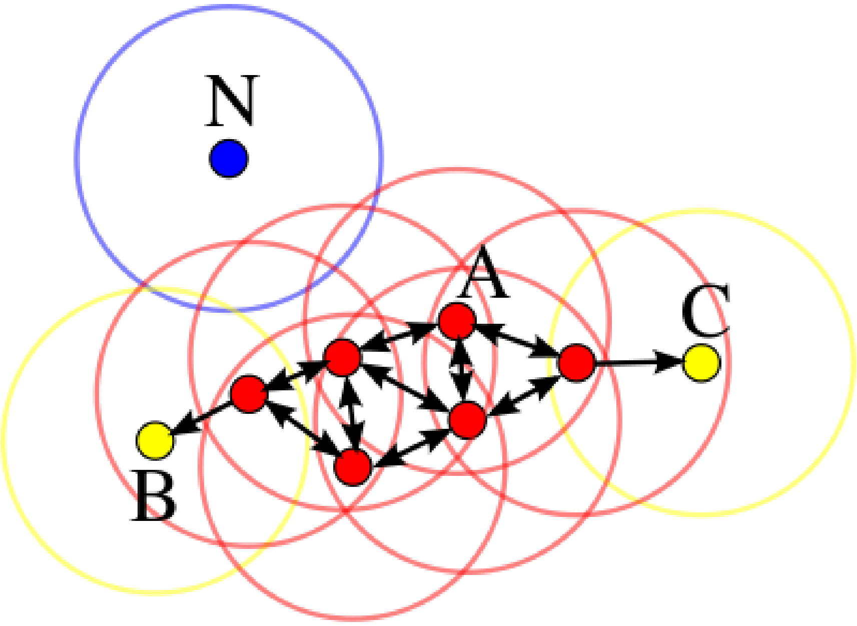
Dbscan (Density Based Spatial Clustering of Applications with Noise)

In machine learning and data analytics clustering methods are useful tools that help us visualize and understand data better. Relationships between features, trends and populations in a data set can be graphically represented via clustering methods like dbscan, and can also be applied to detect outliers in nonparametric distributions in many dimensions.

Dbscan is a density based clustering algorithm, it is focused on finding neighbors by density (MinPts) on an 'n-dimensional sphere' with radius ε. A cluster can be defined as the maximal set of 'density connected points' in the feature space.

Dbscan then defines different classes of points:

- ***Core point:*** **A** is a core point if its neighborhood (defined by ε) contains at least the same number or more points than the parameter MinPts.
- ***Border point***: **C** is a border point that lies in a cluster and its neighborhood does not contain more points than MinPts, but it is still '*density reachable'* by other points in the cluster.
- ***Outlier***: **N** is an outlier point that lies in no cluster and it is not '*density reachable'* nor '*density connected'* to any other point. Thus this point will have "his own cluster".



- If **A** is a core point, it forms a cluster with all the points that are reachable from it. **A** point **Q** is reachable from **P** if there is a path $\boldsymbol{p_1, \ldots, p_n}$ with $\boldsymbol{p_1 = p}$ and $\boldsymbol{p_n = q}$, where each $\boldsymbol{p_{i+1}}$ is directly reachable from $\boldsymbol{p_i}$ (all the points on the path must be core points, with the possible exception of $\boldsymbol{q}$).
- Reachability is a non-symmetric relation since, by definition, no point may be reachable from a non-core point, regardless of distance (so a non-core point may be reachable, but nothing can be reached from it!). Therefore a further notion of *connectedness* is needed to formally define the extent of the clusters found by this algorithm.
- Two points *p* and *q* are density-connected if there is a point ***o*** such that both ***p*** and ***q*** are density-reachable from ***o***. *Density-connectedness* is symmetric.

A cluster satisfies two properties:

- All points within the cluster are mutually density-connected.

- If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

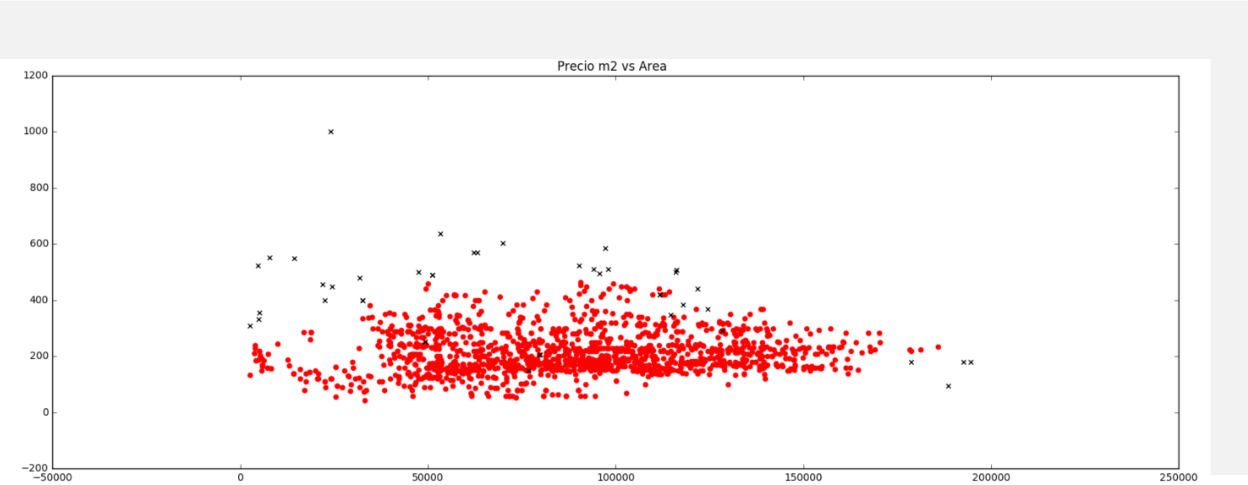
Steps to be followed:

- **Before starting DBSCAN , important step is scaling the data, since the radius ε will define the neighborhoods along with MinPts.**
- After scaling the feature space, is time to choose the spatial metric on which dbscan will perform the clustering. **The metric must be chosen depending on the problem, an euclidean metric works well for 2 or 3 dimensions, the manhattan metric can also be useful when dealing with higher dimensional feature spaces 4 or more dimensions**.
- Then, the parameter eps (ε) must be chosen accordingly to perform clustering. If ε is too big many points will be density connected, if its too small the clustering will result in many meaningless clusters. A good approach is to try values ranging from 0.25 to 0.75.

Dbscan is also sensitive to the MinPts parameter, tuning it will completely depend on the problem at hand.

The complexity of dbscan is of O(*n* log *n*), it and effective method with medium sized data sets. Dbscan estimates the number of clusters by itself, there is no need to specify the number of desired clusters, it is an unsupervised machine learning model.

Outliers (noise) will be assigned to the -1 cluster. After tagging those instances, they can be removed or analyzed.



Real world application of DBSCAN in housing prices (red:normal, black: outliers)

Z-Score pros:

- It is a very effective method if you can describe the values in the feature space with a gaussian distribution. (Parametric)
- The implementation is very easy using pandas and scipy.stats libraries.

Z-Score cons:

- It is only convenient to use in a low dimensional feature space, in a small to medium sized dataset.
- Is not recommended when distributions can not be assumed to be parametric.

Dbscan pros:

- It is a super effective method when the distribution of values in the feature space can not be assumed.
- Works well if the feature space for searching outliers is multidimensional (ie. 3 or more dimensions)
- Sci-kit learn’s implementation is easy to use and the documentation is superb.
- Visualizing the results is easy and the method itself is very intuitive.

Dbscan cons:

- The values in the feature space need to be scaled accordingly.
- Selecting the optimal parameters eps, MinPts and metric can be difficult since it is very sensitive to any of the three params.
- It is an unsupervised model and needs to be re-calibrated each time a new batch of data is analyzed.
    - It can predict once calibrated but is strongly not recommended.

Proximity based methods:

Supervised method

We model an outlier detection as a classification problem. Samples examined by domain experts used for training & testing.

Challenges:

- Classes are unbalanced. That is, the population of outliers is typically much smaller than that of normal objects. Methods for handling unbalanced classes can be used, **such as oversampling**.
- Catch as many outliers as possible, **i.e., recall is more important than accuracy** (i.e., not mislabeling normal objects as outliers)

Unsupervised method

- In some application scenarios, objects labeled as "normal" or "outlier" are not available. Thus, an unsupervised learning method has to be used.
- Unsupervised outlier detection methods make an implicit assumption: The normal objects are somewhat "clustered."
- In other words, an unsupervised outlier detection method expects that normal objects follow a pattern far more frequently than outliers.

Challenges:

- Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
- In case if normal activities are diverse and do not fall into high-quality clusters unsupervised methods may have a high false positive rate and may let many actual outliers be undetected.

The latest unsupervised methods developed smart ideas to tackle outliers directly without explicitly detecting clusters.

Semi-Supervised Methods

In many applications, although obtaining some labeled examples is feasible, the number of such labeled examples is often small. If some labeled normal objects are available:

- Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
- Those not fitting the model of normal objects are detected as outliers

Challenges

- If only some labeled outliers are available, a small number of labeled outliers may not cover the possible outliers well.