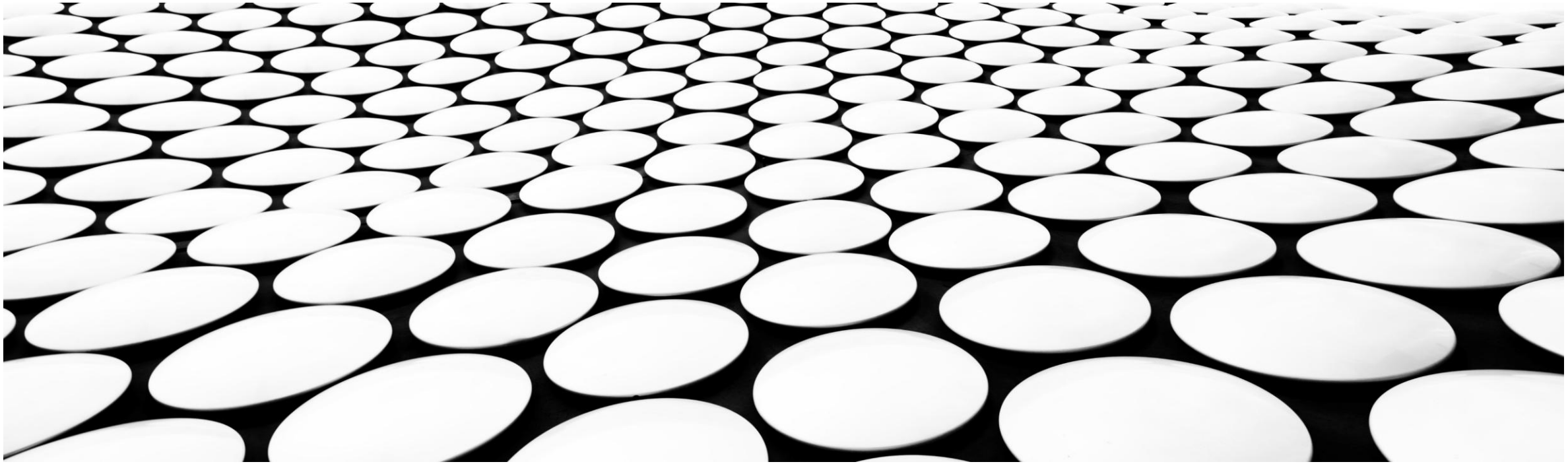
DATA CLEANING

A focus on outliers



DATA CLEANING

- Data wrangling
- Data quality
- Data cleaning
  - Duplicates
  - Invalid data
  - Outliers

Part 1

Part 2

I'll present missing data and data bias later (week 6)

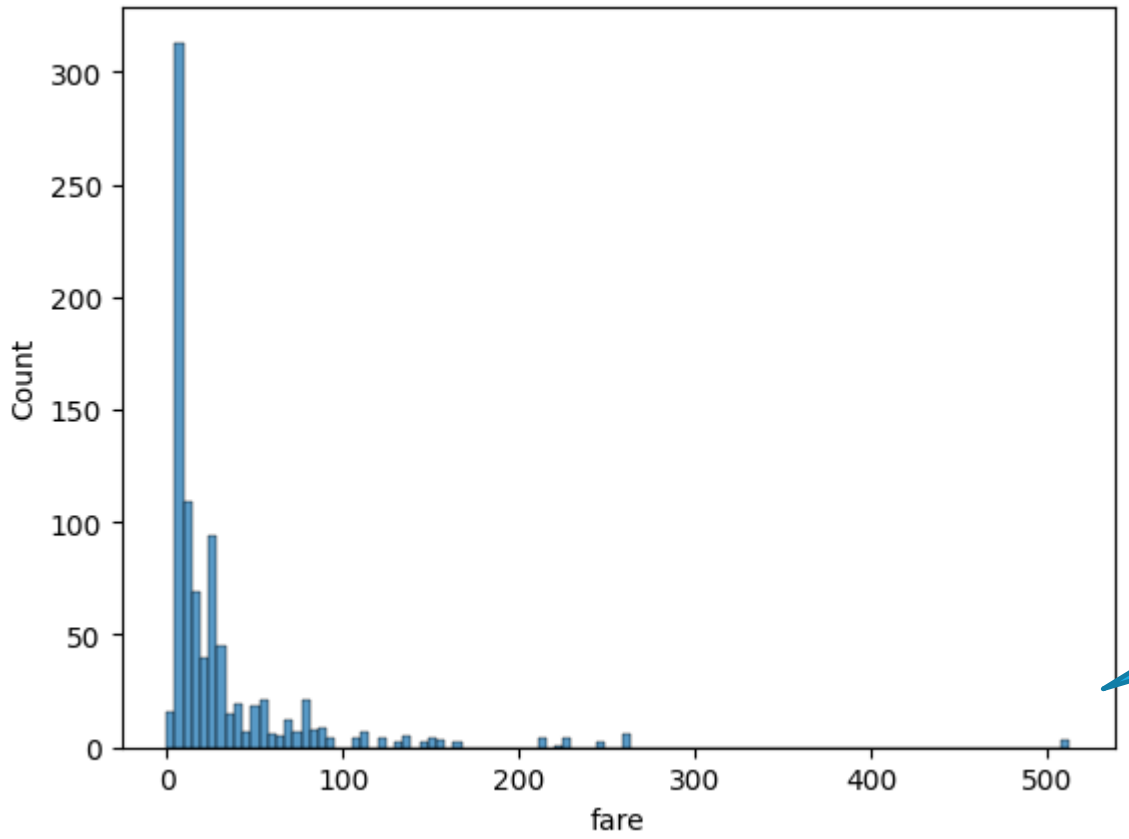
Outliers

Overview

Outliers

Outliers are data points that are significantly different from the majority and can distort averages, correlations, or model performance.

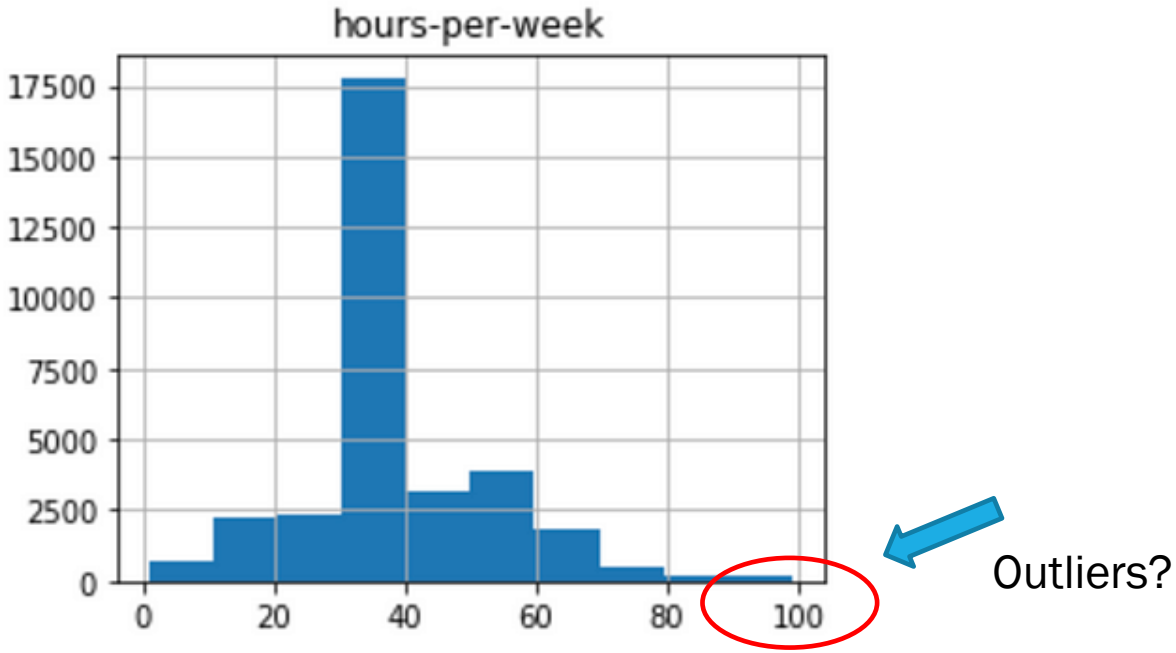
Titanic



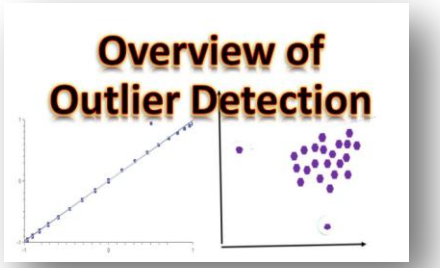
Perhaps an outlier? Or the penthouse?

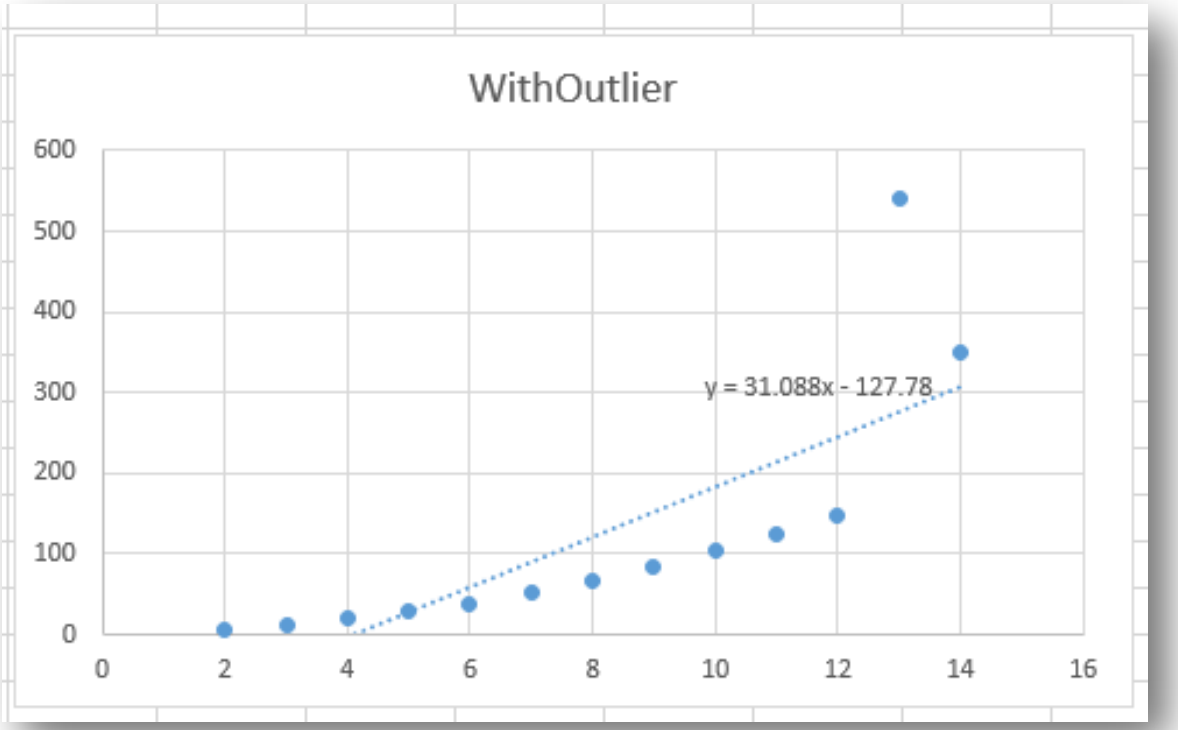
Census

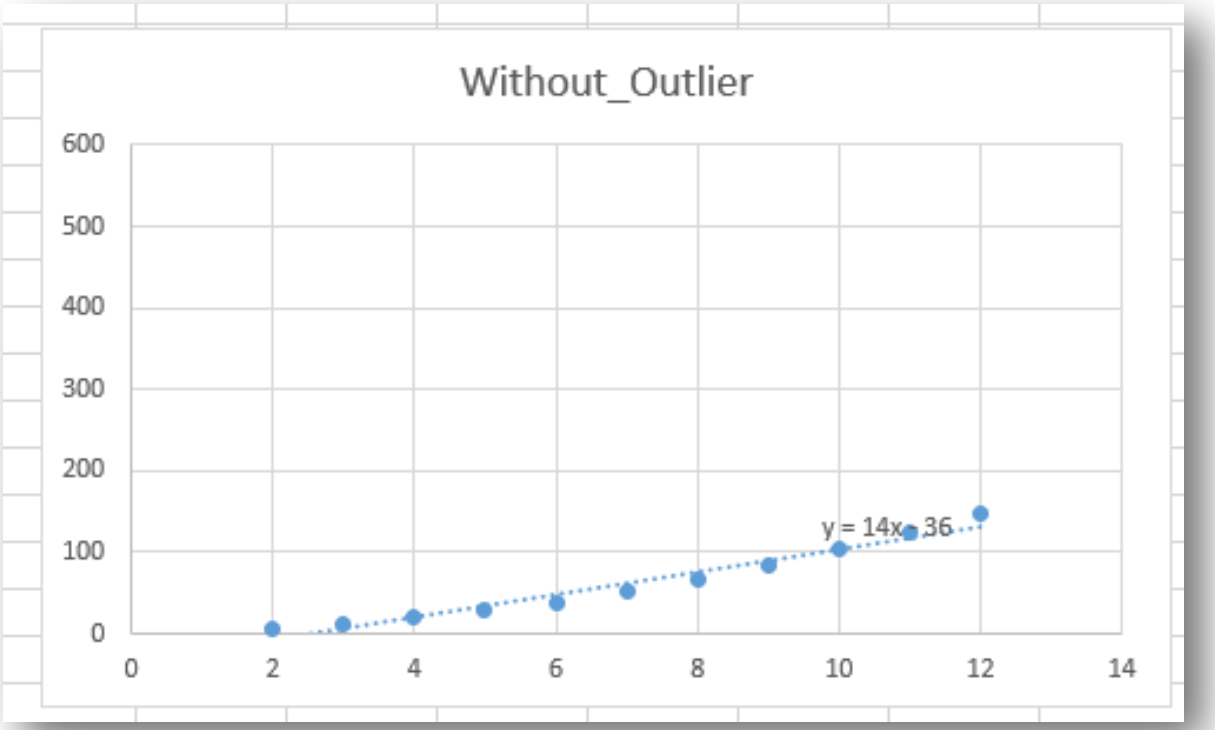
	hours-per-week
count	32561.000000
mean	40.437456
std	12.347429
min	1.000000
25%	40.000000
50%	40.000000
75%	45.000000
max	99.000000



Outliers?

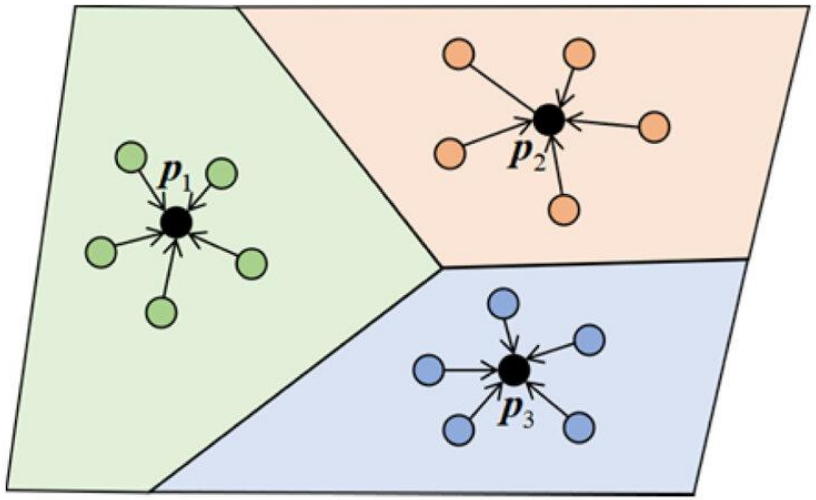

Overview of
Outlier Detection


WithOutlier
600
500
400
300
200
100
0
0
2
4
6
8
10
12
14
16
y = 31.088x - 127.78

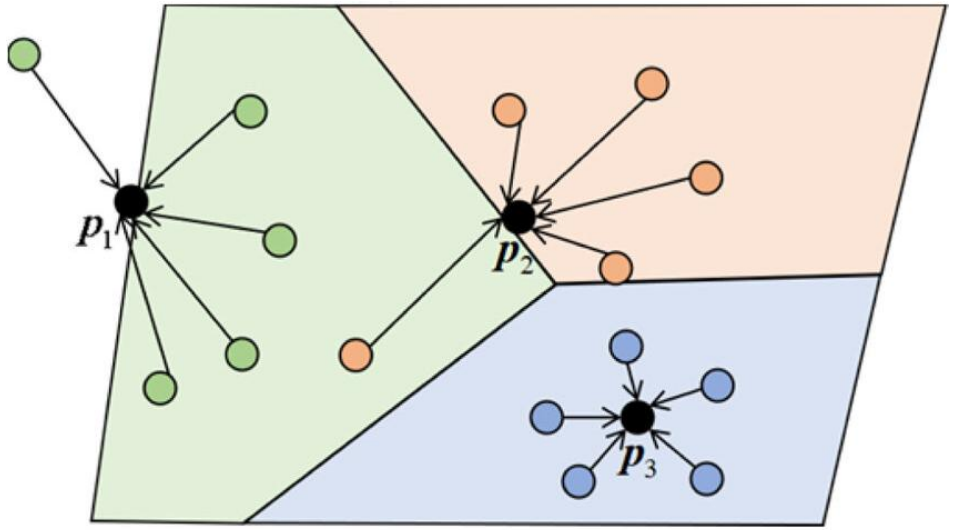

Without_Outlier
600
500
400
300
200
100
0
0
2
4
6
8
10
12
14
y = 14x - 36

Source

Impact on finding centroids for clustering



(a)No outliers



(b)With outliers

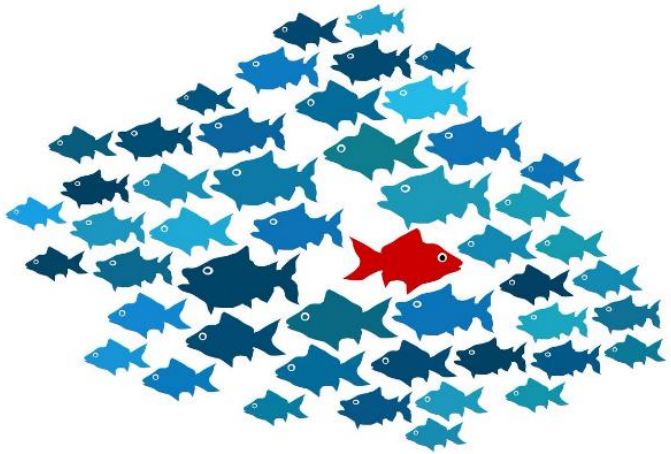
Source

Data cleaning

Detection

Some of the most popular 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)



Detection

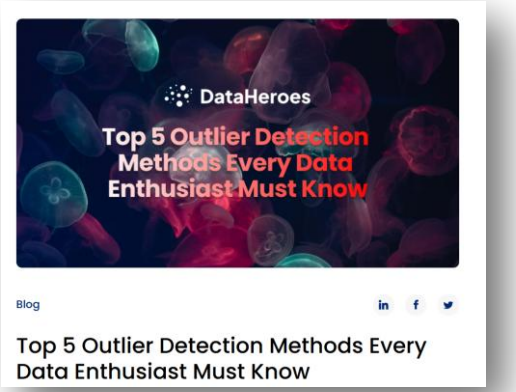
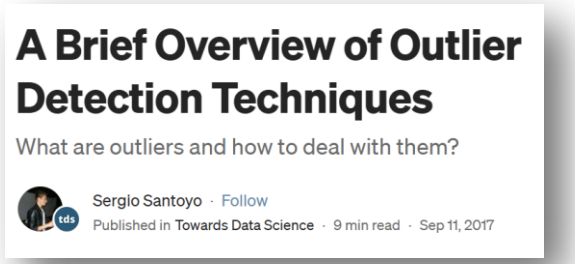
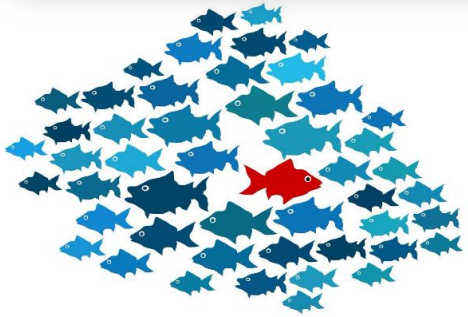


Table of Contents

Source

Understand





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)
- 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) → temperature

Source

Action

How to handle outliers

1) **Remove outliers:** if they are likely data entry errors, could be a good strategy
2) **Transform the data**: change the scale to reduce the impact of outliers, for example, use a log scale
3) **Use robust statistics**: use metrics less sensitive to outliers (e.g. median)
4) **Transform into missing data**: Replace the data by a « missing » value

Ignore the data

Change the view

Change the analysis

Send it to the « missing data » module!

Data cleaning

Knowledge of « normality » or similarity analysis.

Understand data entry process, experiments, measures, etc.

Remove
Transform the view (log scale)
Change the type of analysis
Mark as missing (imputation module)

Detection

Four approaches

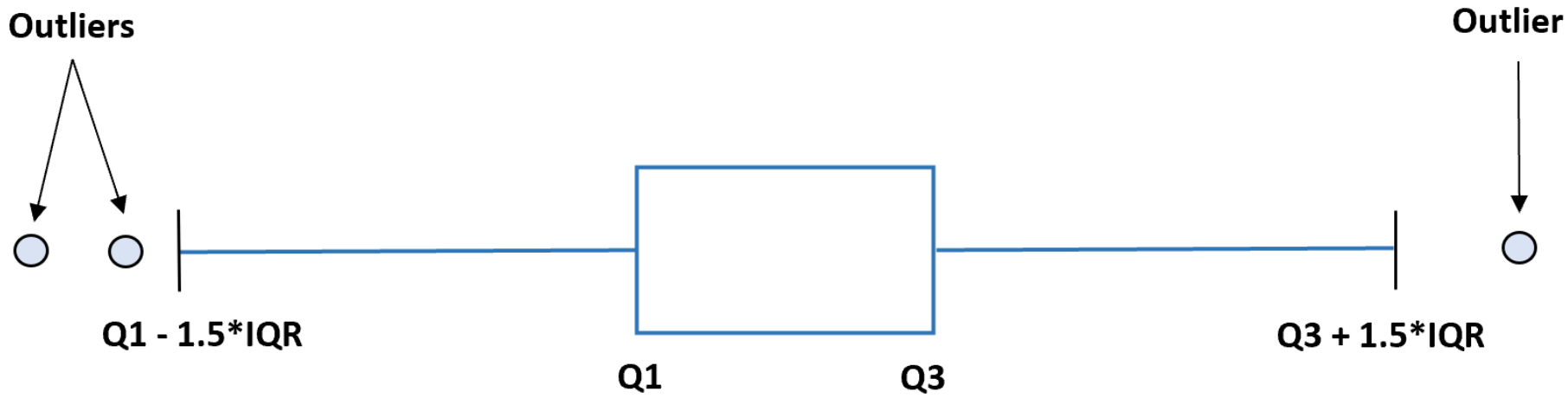
1) Interquartile Range
2) Z-score
3) K-NN
4) Local Outlier Factor

1) Interquartile Range
2) Z-score
3) K-NN
4) Local Outlier Factor

Interquartile Range

The interquartile range, often abbreviated **IQR**, is the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a dataset. It measures the spread of the middle 50% of values.

One popular method is to declare an observation to be an outlier if it has a value 1.5 times greater than the IQR or 1.5 times less than the IQR.



Data
4
5
7
7
8
9
10
12
17
18
19
22
44

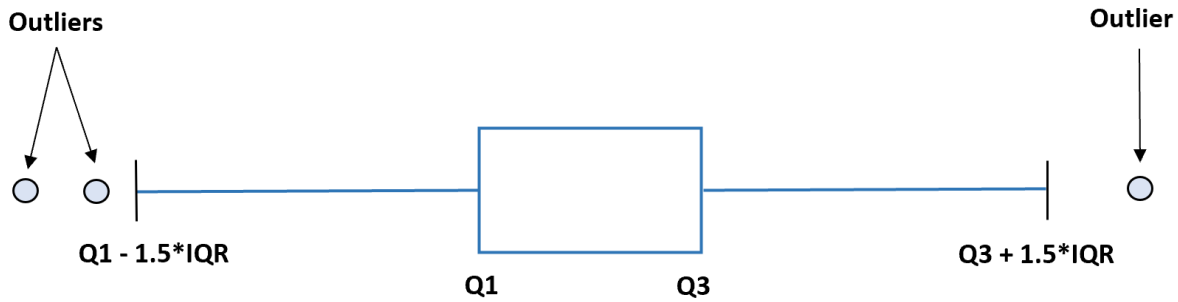
Q1 = 7

Median = 10

Q3 = 18.5

44 → outlier

IQR = 18.5 – 7 = 11.5

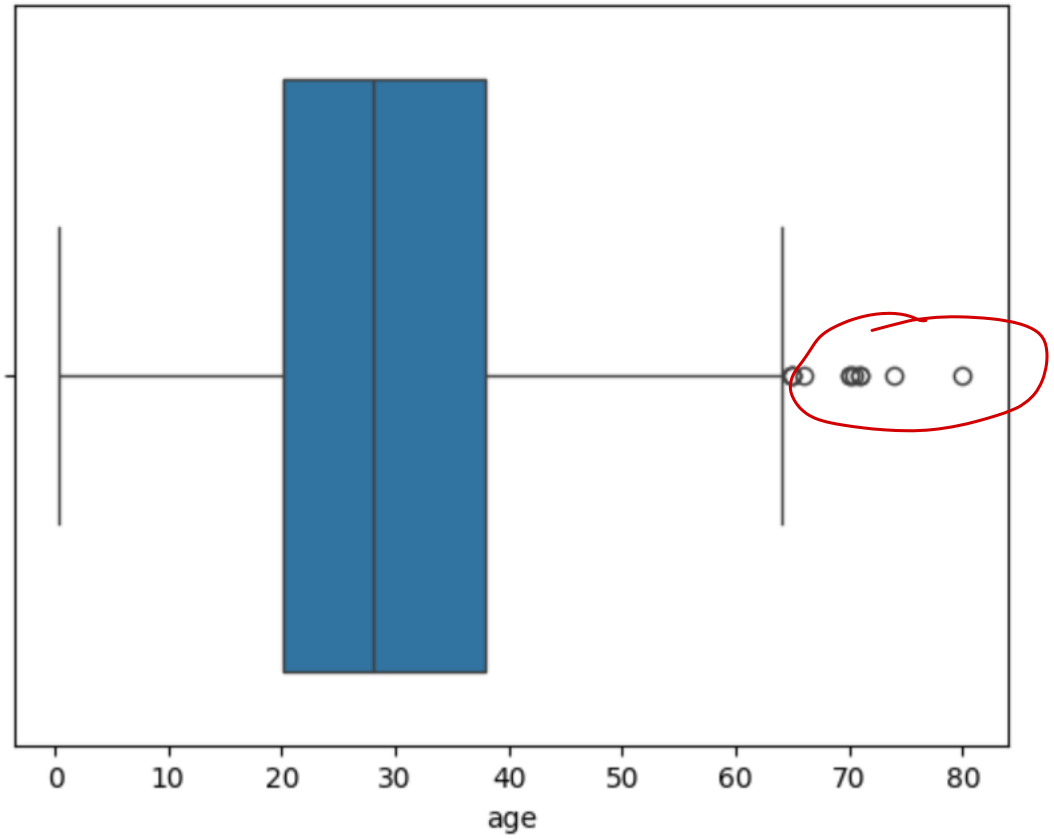


$$Q1 - 1.5*IQR \leq \text{Acceptable range} \leq Q3 + 1.5*IQR$$
$$Q1 - 17.25 \leq \text{Acceptable range} \leq Q3 + 17.25$$
$$-10.25 \leq \text{Acceptable range} \leq 35.75$$

Titanic

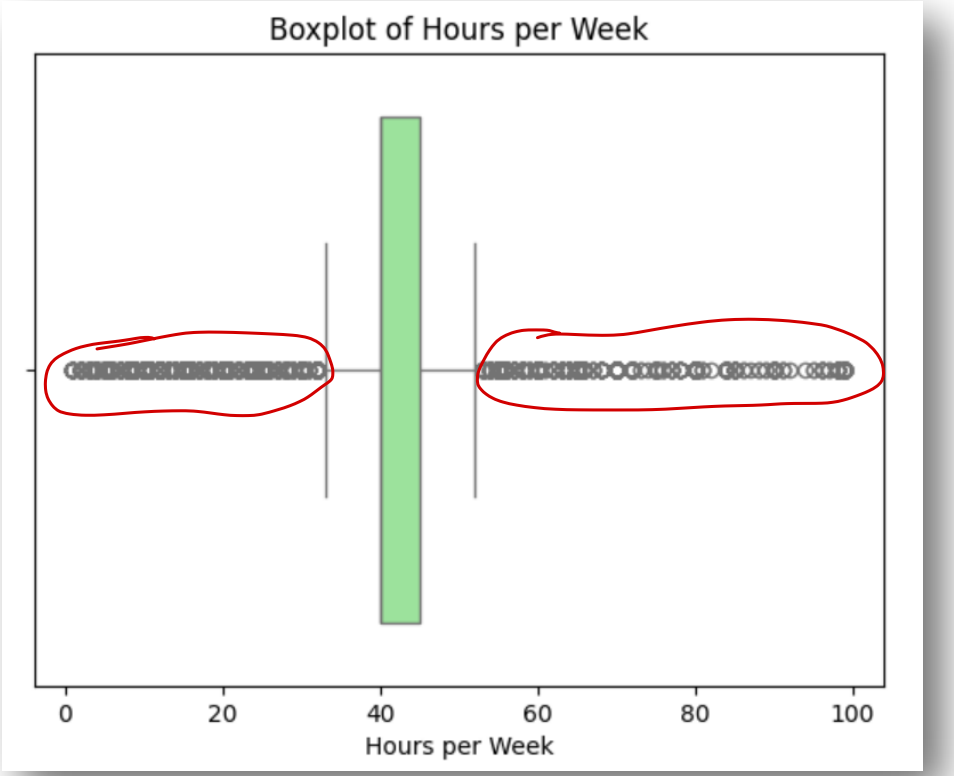
```
titanic[["age"]].describe()
```

	age
count	714.000000
mean	29.699118
std	14.526497
min	0.420000
25%	20.125000
50%	28.000000
75%	38.000000
max	80.000000



Census

	hours-per-week
count	32561.000000
mean	40.437456
std	12.347429
min	1.000000
25%	40.000000
50%	40.000000
75%	45.000000
max	99.000000



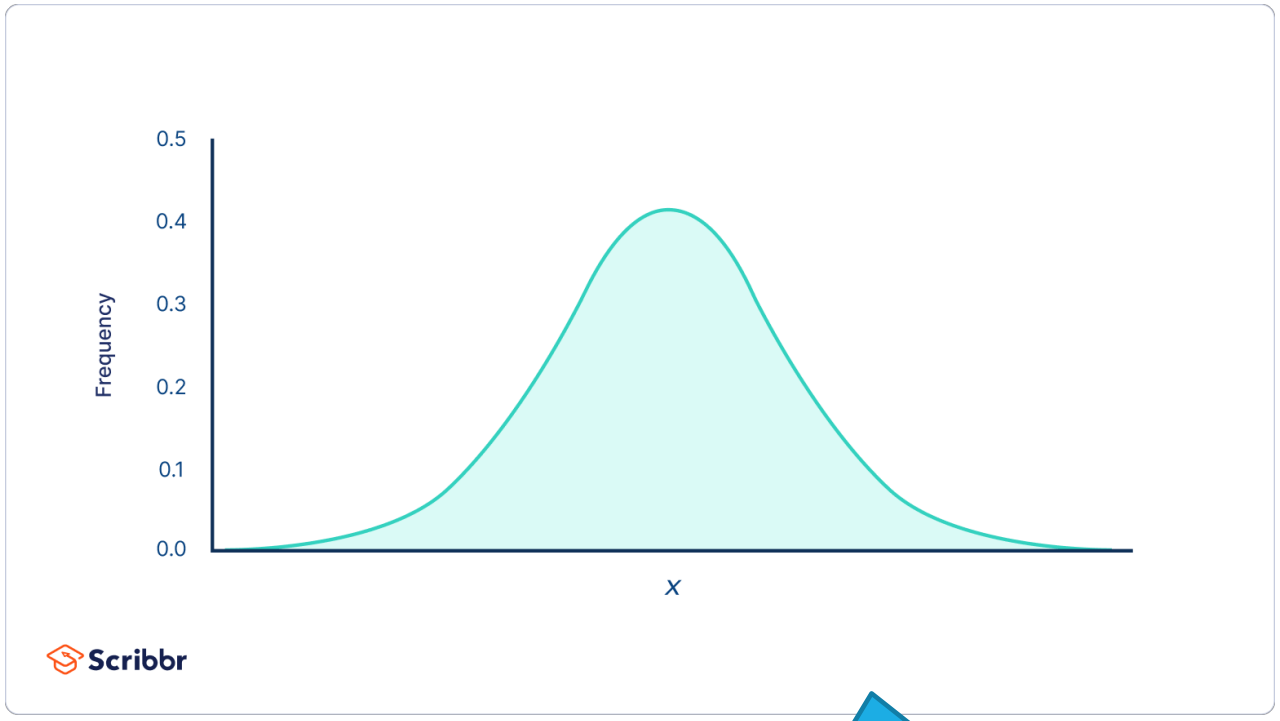
1) Interquartile Range
2) Z-score
3) K-NN
4) Local Outlier Factor

Z-score

The Z-score method is a statistically based approach for outlier detection. It computes the standard score, or Z-score, for each data point.

We then set a threshold for our Z-score, and data points with Z-scores greater than it are considered outliers.

An important assumption made by the Z-score method is that **your data is normally distributed,** making it especially useful for datasets with symmetrical patterns around the mean.



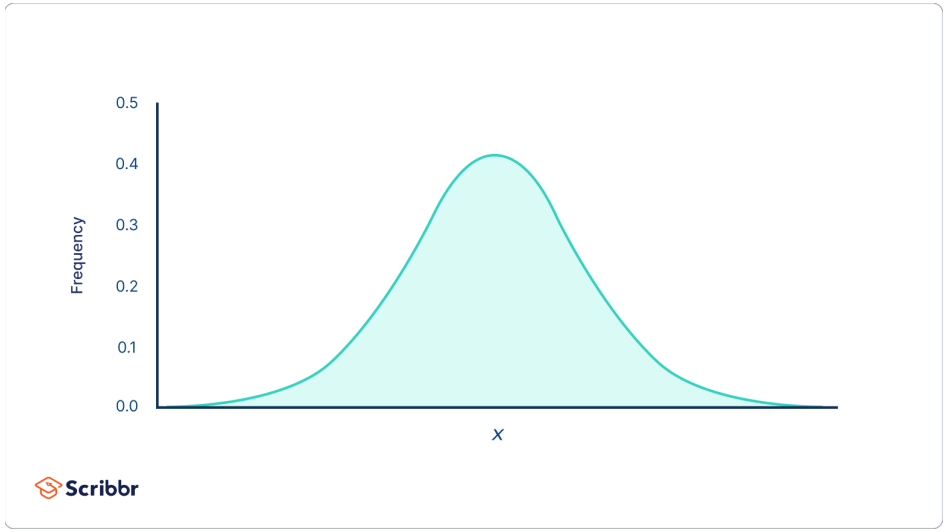
Normal distribution is often a good representation of variations in natural phenomena

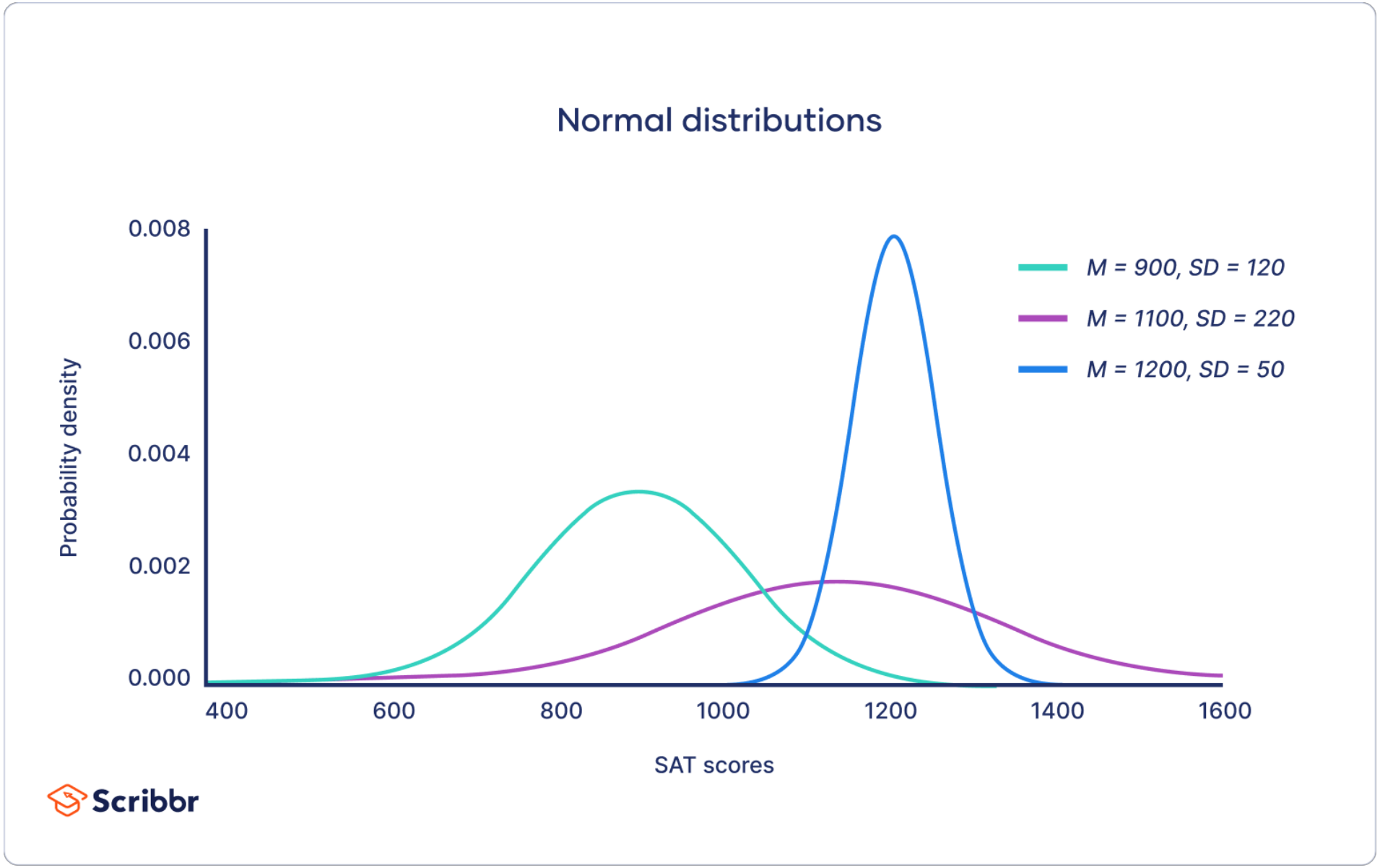
What are the properties of normal distributions?

Normal distributions have key characteristics that are easy to spot in graphs:

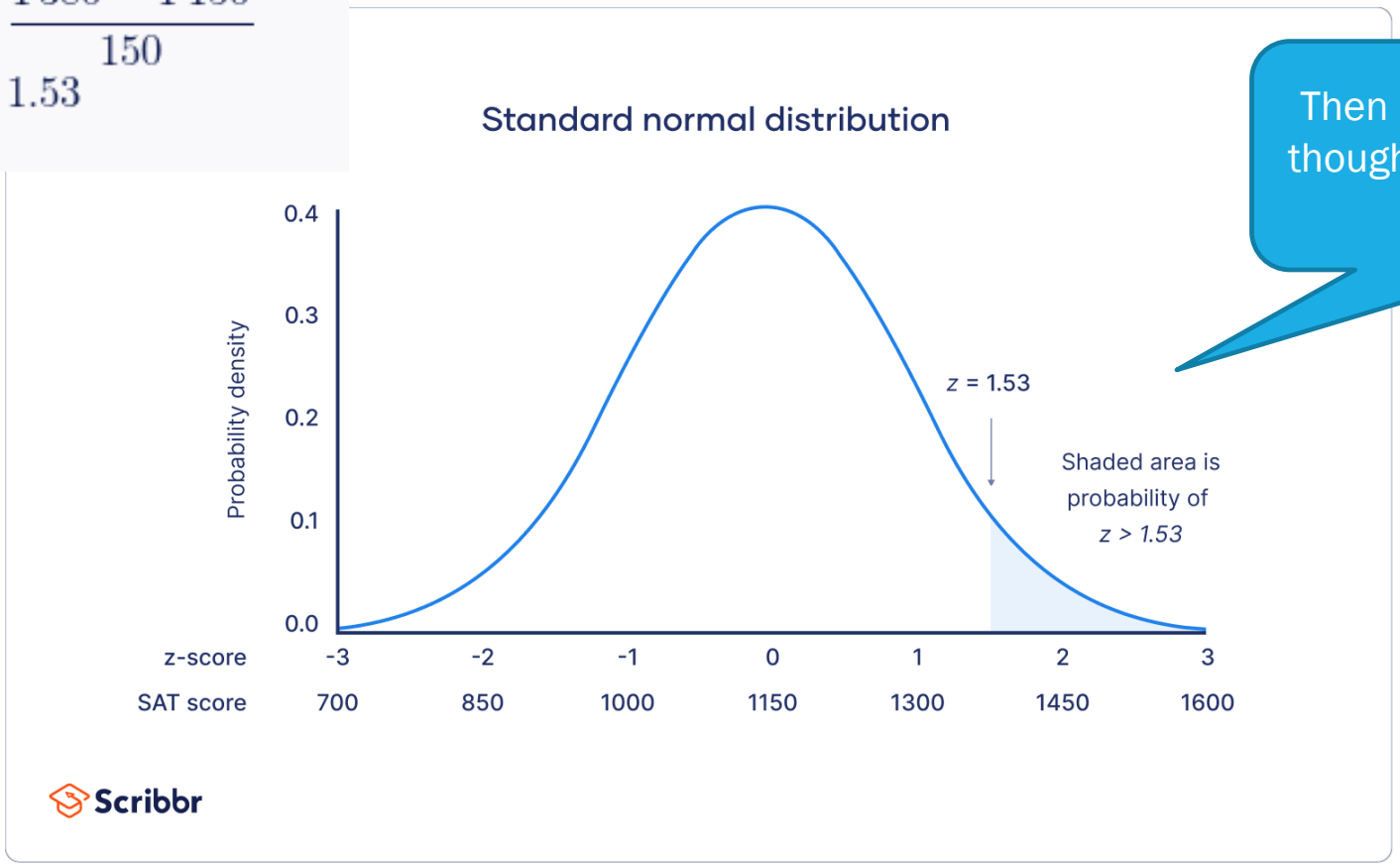
- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

Normal probability density formula	Explanation
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	• $f(x)$ = probability • x = value of the variable • μ = mean • σ = standard deviation • σ^2 = variance



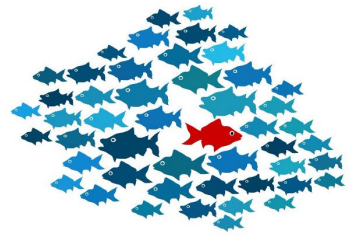

Normal distributions
0.008
0.006
0.004
0.002
0.000
Probability density
400
600
800
1000
1200
1400
1600
SAT scores
M = 900, SD = 120
M = 1100, SD = 220
M = 1200, SD = 50
Scribbr

Formula	Calculation
$z = \frac{x - \mu}{\sigma}$	$z = \frac{1\,380 - 1\,150}{150}$ $z = 1.53$



Then z-score obtained can be thought as number of std away from the mean

Coding example



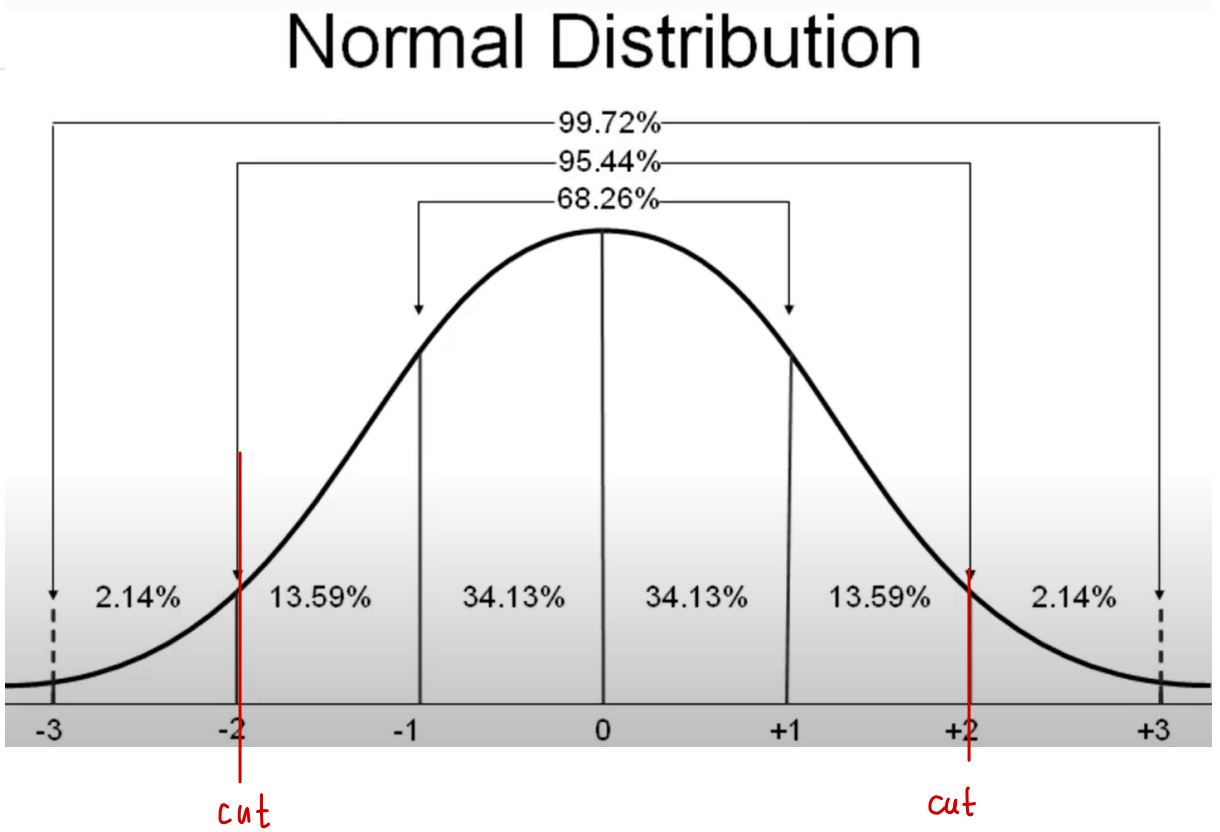
```
from sklearn.datasets import load_breast_cancer
from scipy import stats

threshold = 2.5
df = load_breast_cancer(as_frame=True).data
z_scores = stats.zscore(df)
outliers = df[abs(z_scores) > threshold]
```

2.5

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

Threshold is field dependent



If I ask you to toss the coin 100 times and record how many times you get heads:

- If you get heads 50% of the time, you're okay with the results.
- If you get heads 40% of the time, you're still okay, right?
- If you get heads 10% of the time, you might say, "Umm… okay," but start to have doubts.

- But if you get heads only 5% of the time, this is when you get frustrated and say, "The coin is not fair!" — and you stop!

An interesting intuition (not scientific!) for a tolerance to « abnormality »

Advantages

1. Ease of implementation
2. Assumes that the data is distributed normally, which is a widely applicable assumption for situations in the real world.
3. Offers a numerical assessment of the extremeness of each outlier based on standard deviations.

Disadvantages

1. If your data is not normally distributed, Z-score will not be effective for detecting outliers
2. It may be influenced by the presence of other outliers in the dataset.
3. Depending on the dataset and context, the threshold value selection has to be done carefully.

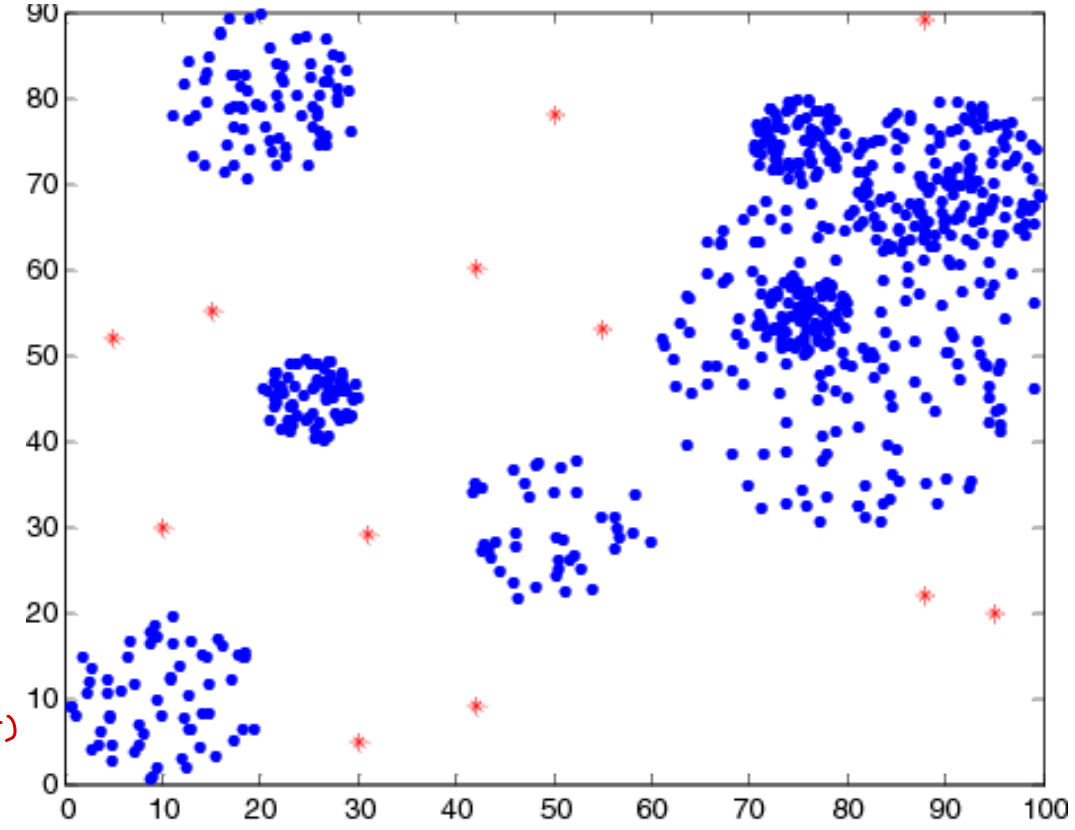
↳ needs domain knowledge

1) Interquartile Range
2) Z-score
3) K-NN
4) Local Outlier Factor

K-Nearest Neighbors

K Nearest Neighbors

Detect points that are *further away* from their neighbours then what is expected from the data distribution.

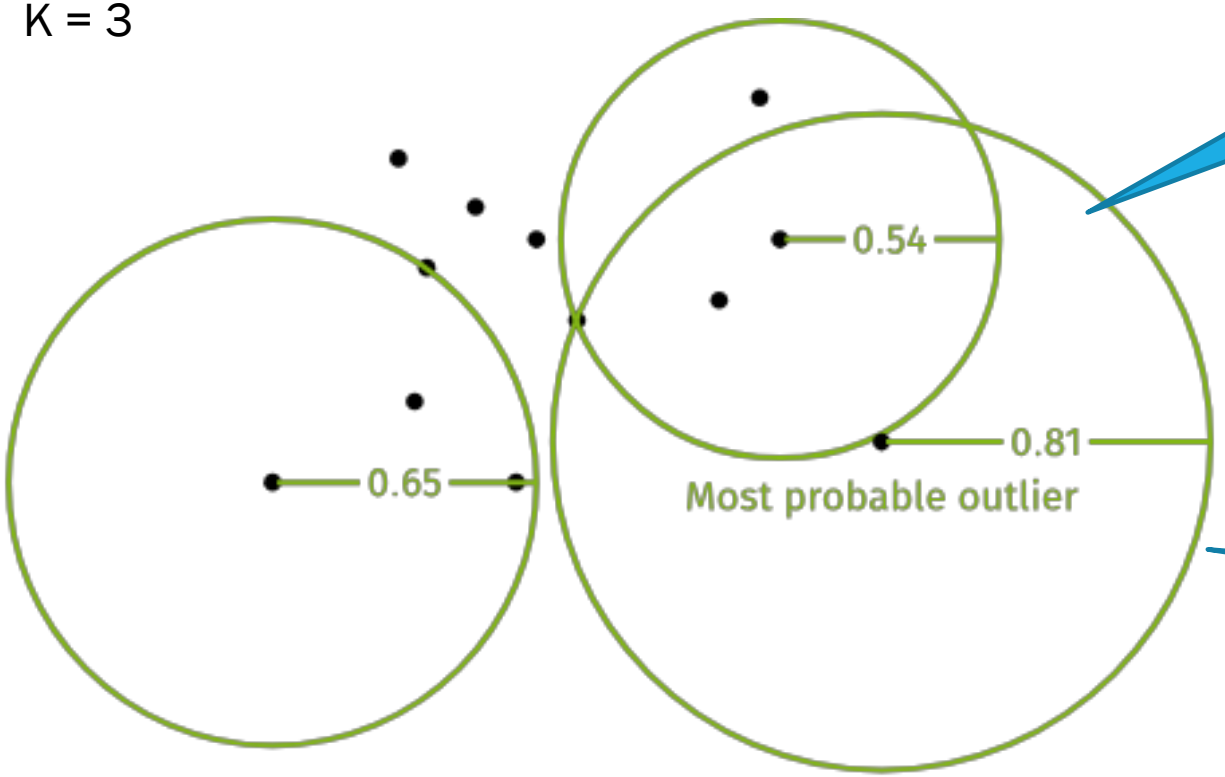


High level algorithm

1) Select k (number of neighbours to consider)
2) Calculate for each point, its distance to its k'th nearest neighbor (called kDistance) distance to 5th nearest neighbour (k=5)
3) Set a threshold (acceptable kDistance) can be tricky
4) Remove points having a kDistance above the threshold

K-Distance

K = 3



Do the calculation for each data point

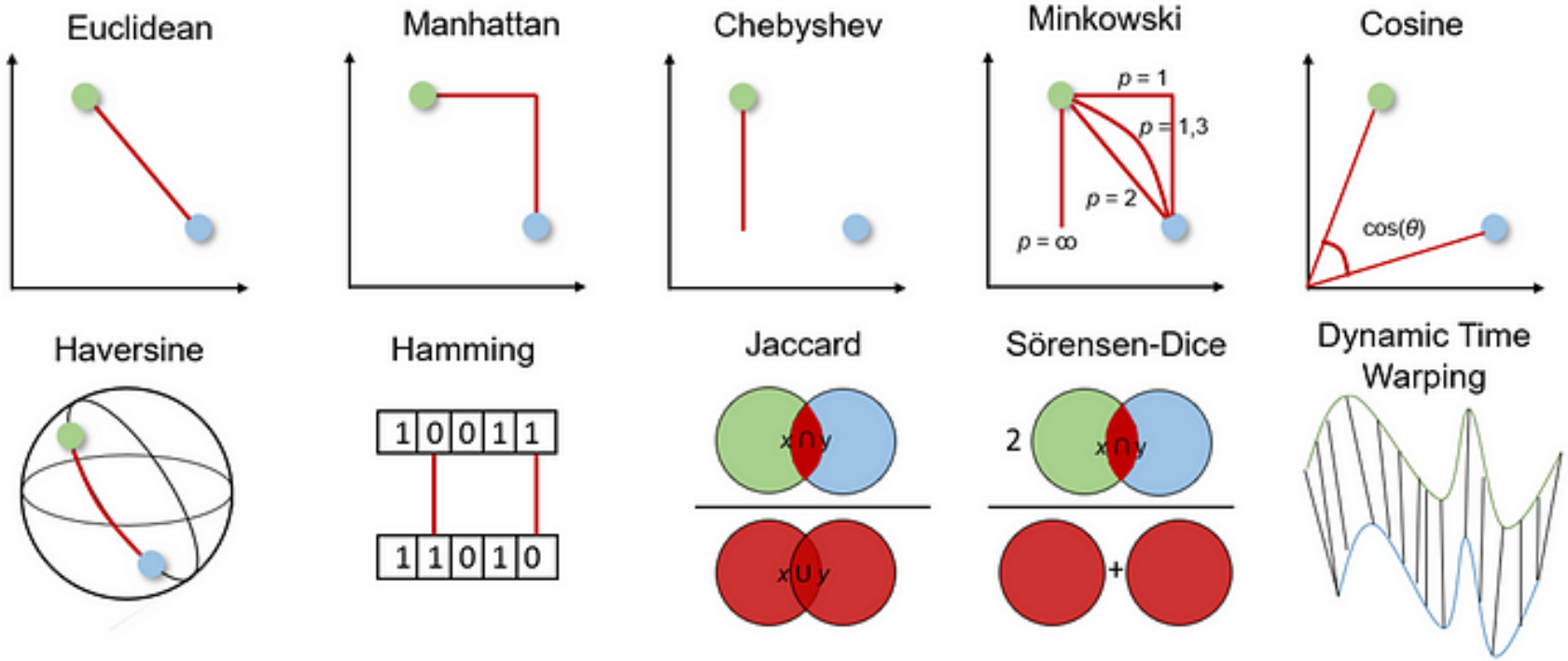
Any alternative to Euclidian Distance?

How to set a threshold?

A brief introduction to Distance Measures

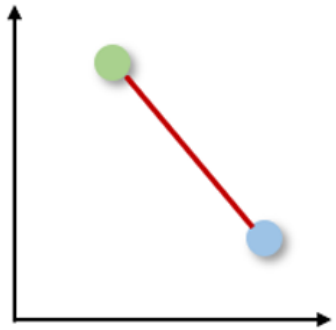
10 distance measures for machine learning you should have heard of

Jonte Dancker · Follow
9 min read · Oct 25, 2022



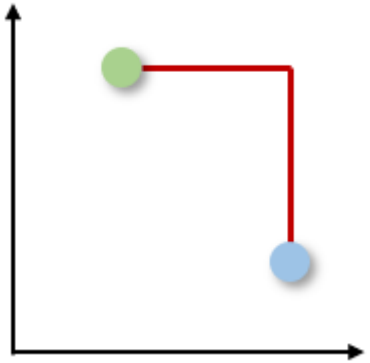
Source

Euclidean



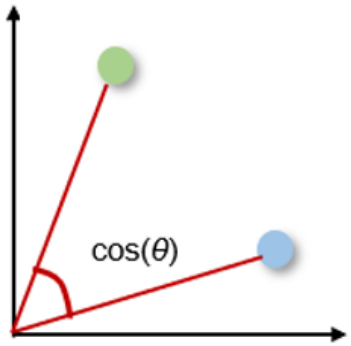
$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan

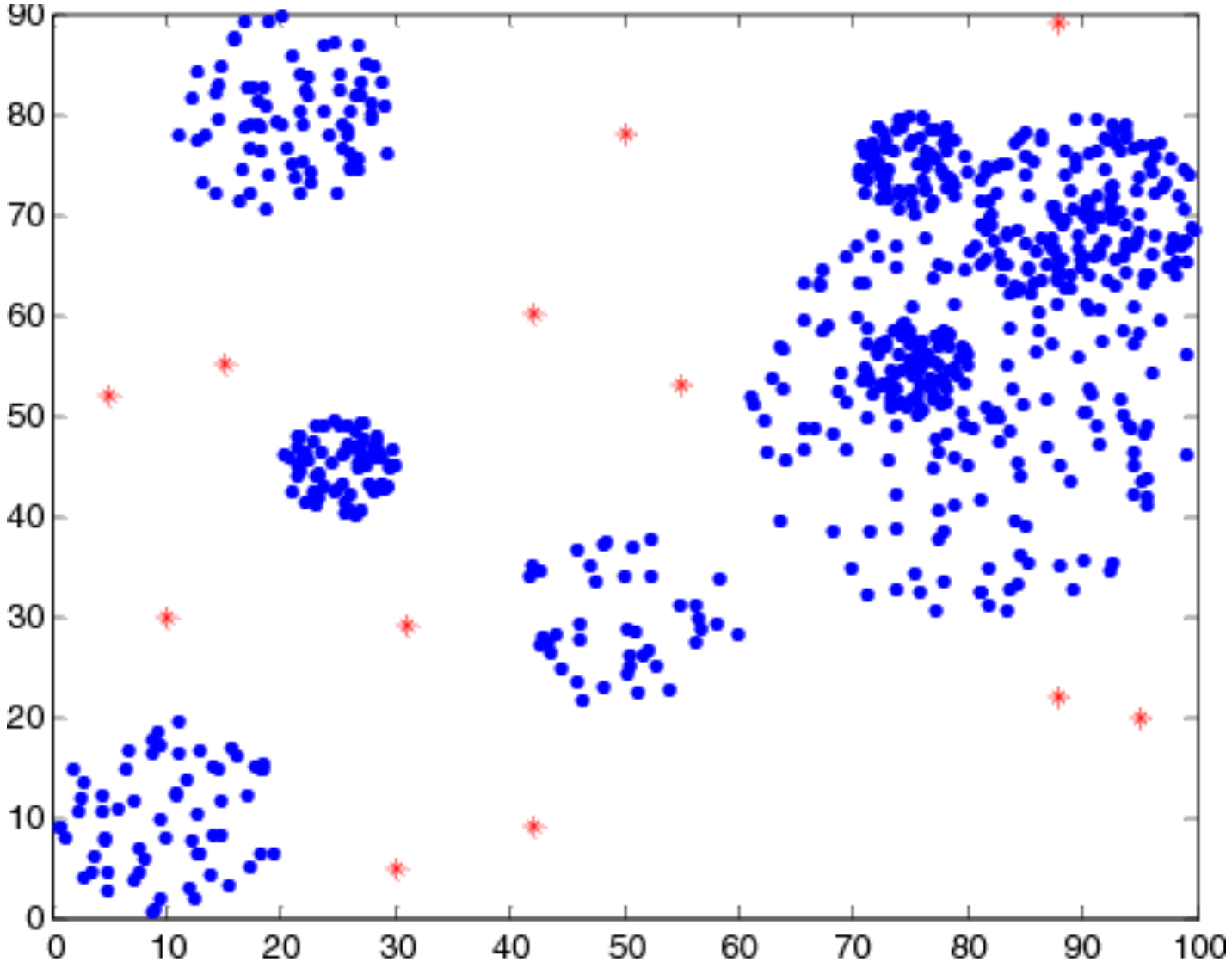


$$d = \sum_{i=1}^{n} (x_i - y_i)$$

Cosine



$$d = \cos(\alpha) = \frac{x\,y}{\|x\|\,\|y\|}$$



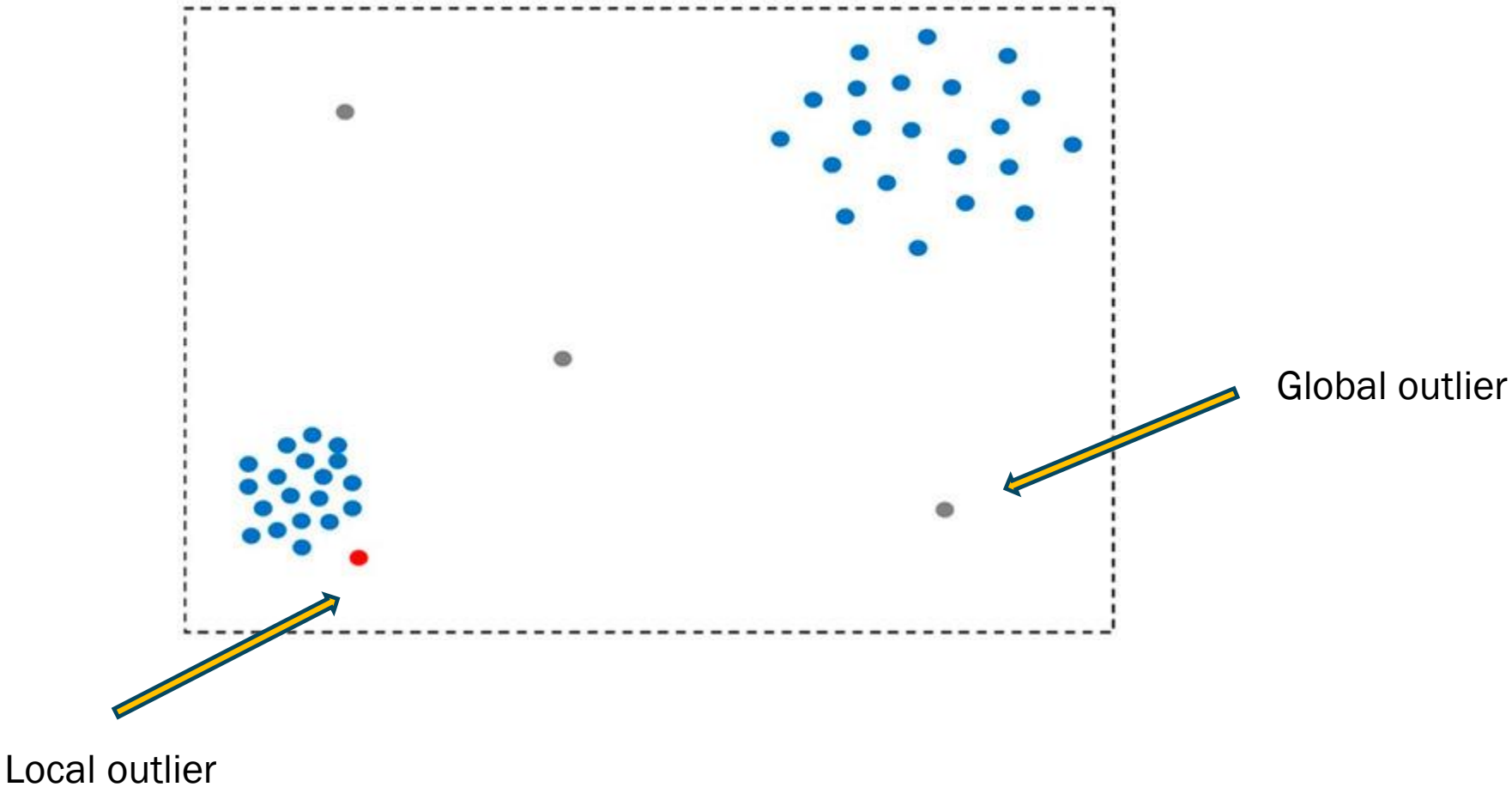
Threshold

Can gather the distribution of k-distances and use z-score.

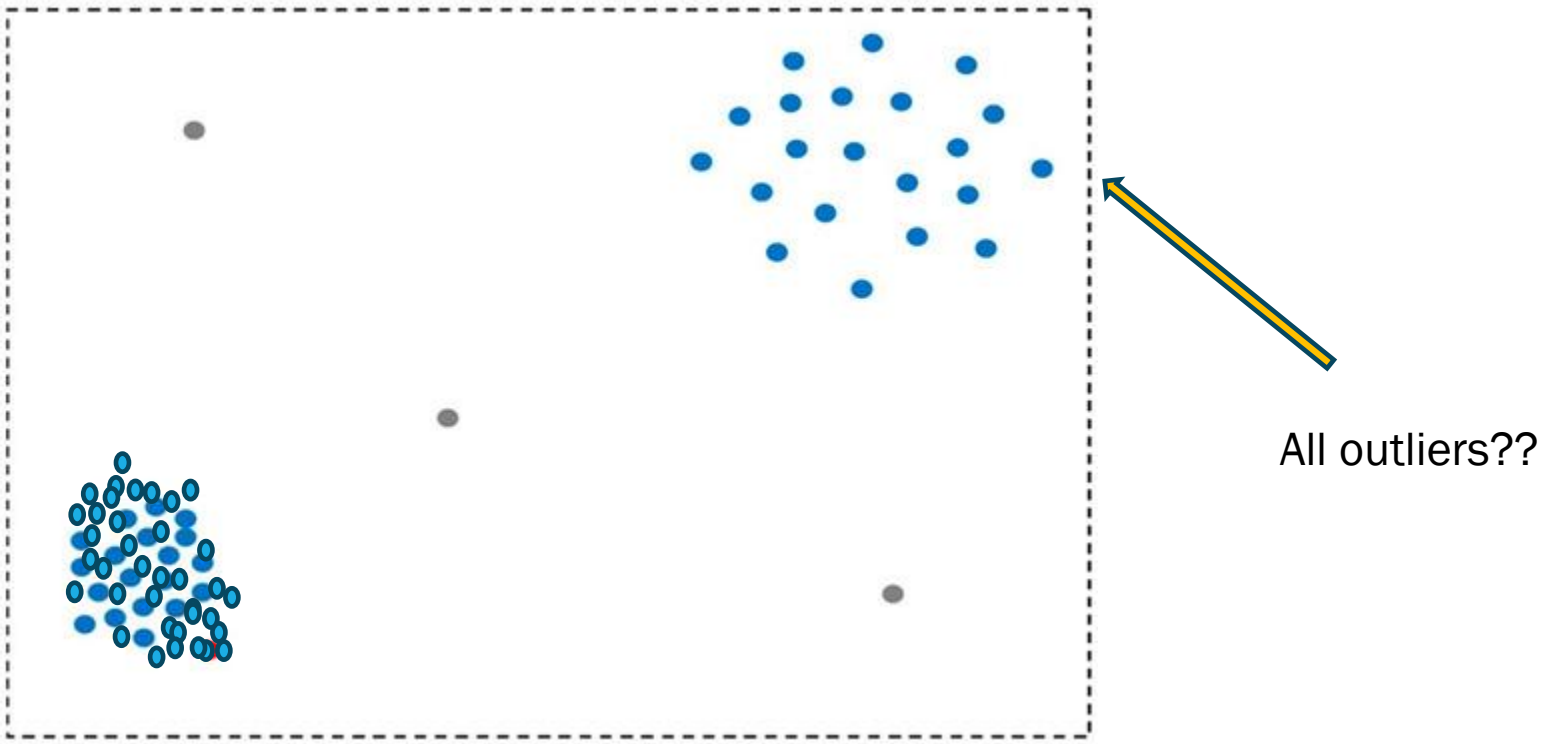
Example:
K=10
For each point, evaluate the distance to it's 10th closest neighbour.

Gather all those k-distances and look at their distributions.

Limitations of kNN

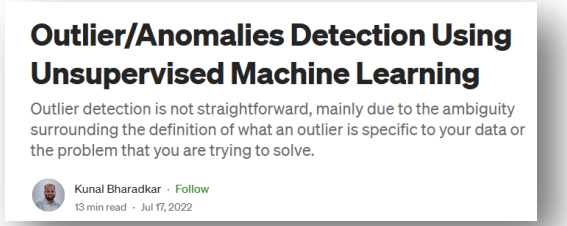


Limitations of kNN



1) Interquartile Range
2) Z-score
3) K-NN
4) Local Outlier Factor

Local Outlier Factor



It is easier to illustrate this with an example: imagine a person standing in line in a small but busy Starbucks, and everyone is pretty much close to each other; then, we can say the person is in a high-density area and, more specifically, **high local density**. If the person decides to wait in their car in the parking lot until the line eases up, they are isolated and in a low-density area, thus being considered an outlier. From the perspective of the people standing in line, who are probably not aware of the person in the car, that person is considered not reachable even though that person in the vehicle can see all of the individuals standing in line. So we say that the person in the car is not reachable from their perspective. Hence, we sometimes refer to this as **inverse reachability** (how far you are from the neighbors' perspective, not just yours).

Source

- We need a Distance Measure, and set K (e.g. 2).
- We need a notion of Neighbors (e.g. 2)
- We need a notion of Reachability
- We need a notion of Average Reachability from the neighbors
- We need to compare that average reachability to that of the neighbors

2 neighbors

average distances

Reachability from its neighbors

only consider clients

Can I be reached by my neighbors the same way that my neighbors can be reached by their neighbors.?

Coding example

```
from sklearn.datasets import load_breast_cancer
from sklearn.neighbors import LocalOutlierFactor

data = load_breast_cancer(as_frame=True).data
lof = LocalOutlierFactor(n_neighbours=20,contamination=0.1)
```

N_neighbours is the number of neighbours to consider, value of K

Contamination is the proportion of outliers expected used to set the threshold on LOF result

Local Outlier Factor (LOF) — Algorithm for outlier identification

Detection of the anomaly using LOF values

Vaibhav Jayaswal · Follow
Published in Towards Data Science · 5 min read · Aug 30, 2020

Distance used is Manhattan distance (could be something else)

MD(A,B) = MD(B,A) = 1
MD(A,C) = MD(C,A) = 2
MD(A,D) = MD(D,A) = 3
MD(B,C) = MD(C,B) = 1
MD(B,D) = MD(D,B) = 4
MC(C,D) = MD(D,C) = 3

Source

K-Distance Distance between a point and it's Kth nearest neighbour

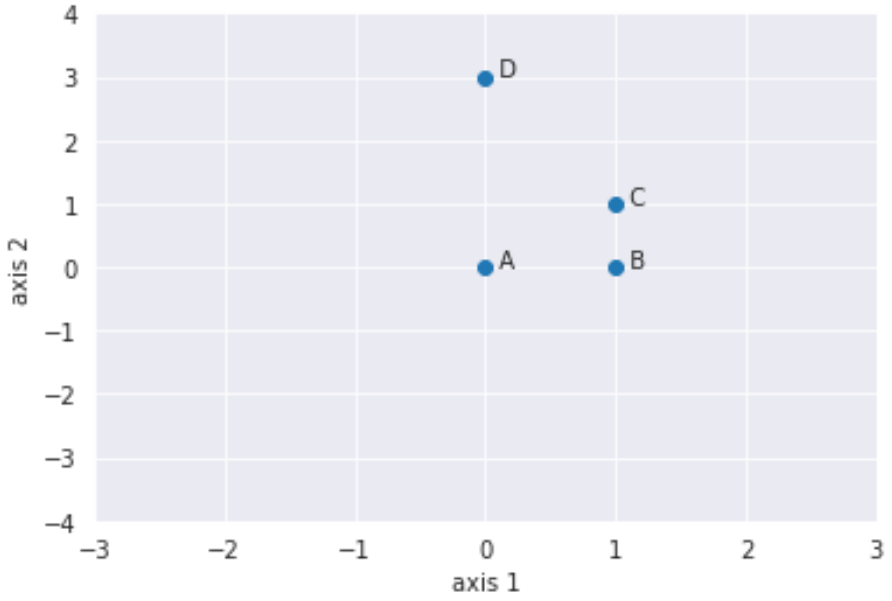
Assume K = 2 : What is the distance to the second closest neighbour?

K_2-distance(A) = MD(A,C) = 2
K_2-distance(B) = MD(B,A) or MD(B,C) = 1
K_2-distance(C) = MD(C,A) = 2
K_2-distance(D) = MD(D,C) or MD(D,A) = 3

K_1-distance(A) = MD(A,B) = 1
K_1-distance(B) = MD(B,A) = 1
K_1-distance(C) = MD(C,B) = 1
K_1-distance(D) = MD(D,A) = 3

MD(A,B) = MD(B,A) = 1
MD(A,C) = MD(C,A) = 2
MD(A,D) = MD(D,A) = 3
MD(B,C) = MD(C,B) = 1
MD(B,D) = MD(D,B) = 4
MC(C,D) = MD(D,C) = 3

K_2-distance(A) = 2
K_2-distance(B) = 1
K_2-distance(C) = 2
K_2-distance(D) = 3



Find the set of all neighbours that are within the K_2-distance established

Size of set

K-neighborhood (A) = {B,C} , ||N2(A)|| =2

K-neighborhood (B) = {A,C}, ||N2(B)|| =2

K-neighborhood (C)= {B,A}, ||N2(C)|| =2

K-neighborhood (D) = {A,C}, ||N2(D)|| =2

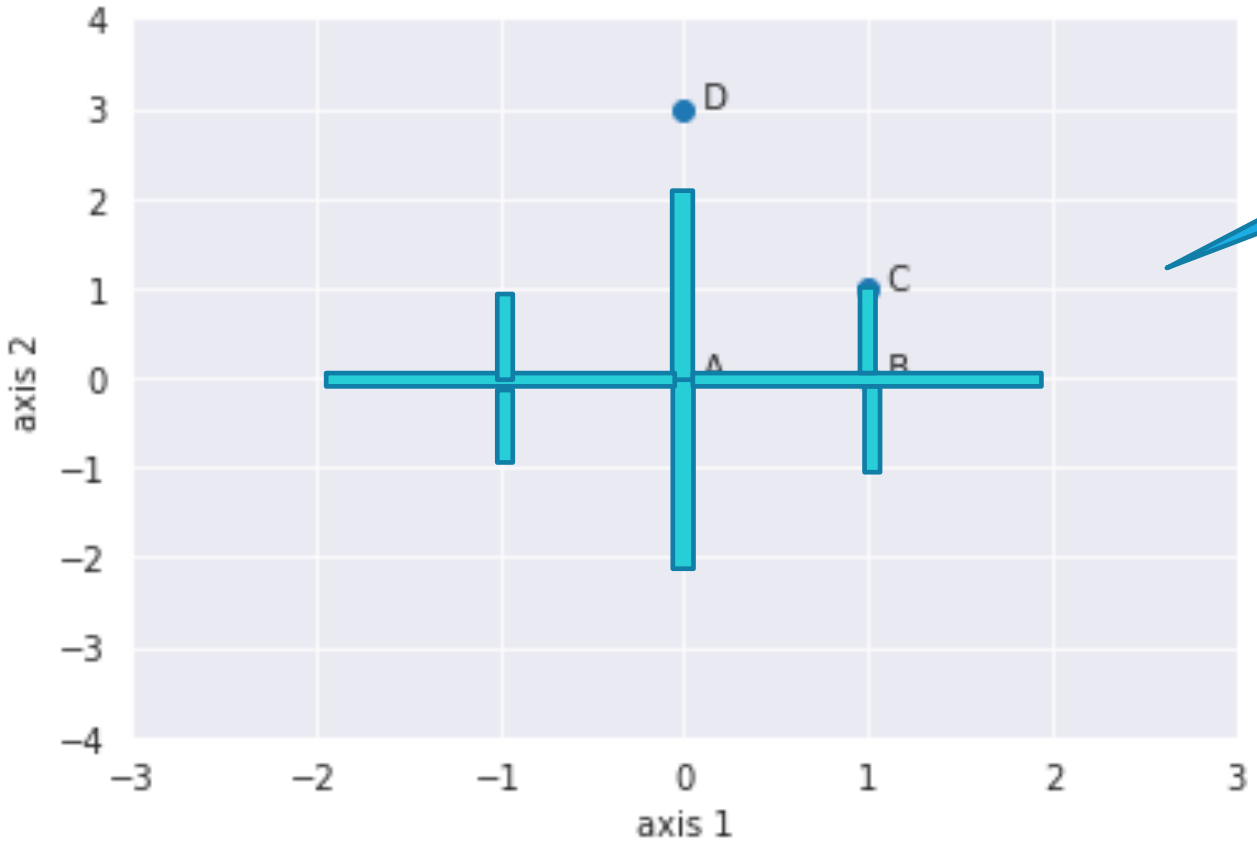
If using K_1-distance

k-neighbourhood (A) = {B} ||N1(A)|| = 1

k-neighbourhood (B) = {A,C} ||N1(B)|| = 2

k-neighbourhood (C) = {B} ||N1(C)|| = 1

k-neighbourhood (D) = {A,C} ||N1(D)|| = 2



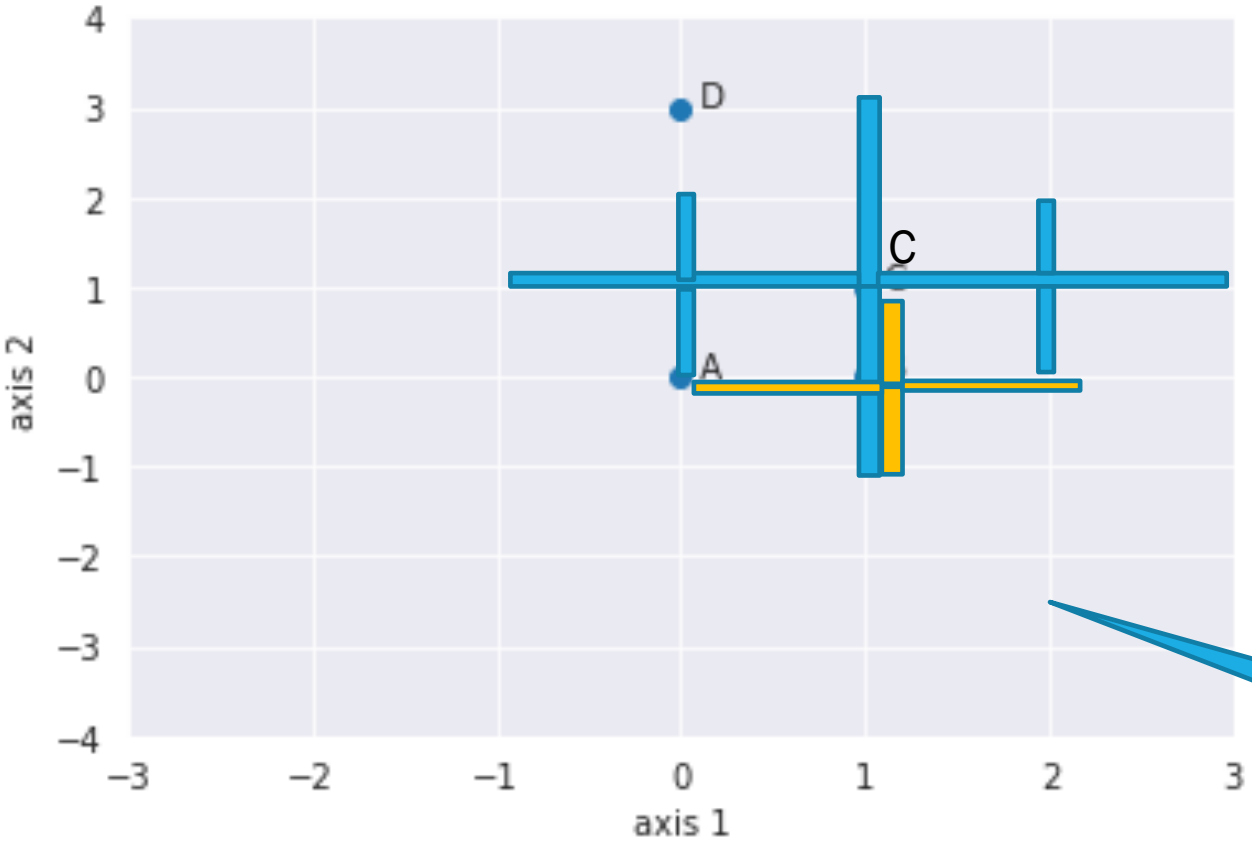
Let's focus on A

K_2-distance(A) = 2

K-neighborhood (A) = {B,C} , ||N2(A)|| =2

Reachability Distance RD

RD(Xi, Xj) = max (Kdistance(Xj) , distance(Xi, Xj)



From A, we know B and C are within K_2-distance(A)

Now, we want to know to what extent A is within the K_2-distance of its neighbours (B and C). Is it « reachable » from its neighbours?

RD(A,B) = max(K_2-distance(B), MD(A,B)) = 1
RD(A,C) = max(K_2-distance(C), MD(A,C)) = 2

From D, A and C are within k_1-distance(B)
RD(D,A) = max (k_1-distance(A), MD(D,A)) = 3
RD(D,C) = max (k_1-distance(C), MD(D,C)) = 3

We look at the neighbours of A

Local Reachability Density

$$LRD_k(A) = \frac{1}{\sum_{X_j \in N_k(A)} \frac{RD(A,X_j)}{\|N_k(A)\|}}$$

Average of reachability to A from all the points that were within K_2-distance(A)

On average, how far (reachability distance) is A from the point of view of its neighbours?
High average (far from neighbors) = low density (LRD).

$$LRD_1(D) = \frac{1}{\frac{3}{2} + \frac{3}{2}} = \frac{1}{3}$$

Local outlier Factor

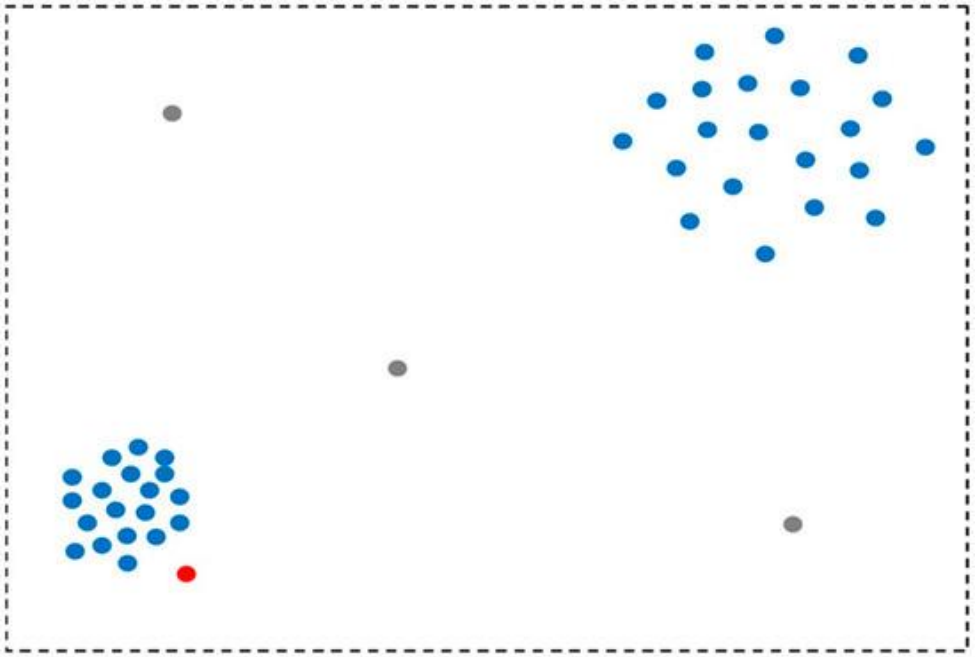
$$LOF_k(A) = \frac{\sum_{Xj \in N_k(A)} LRD_k(Xj)}{||N_k(A)||} \times \frac{1}{LRD_k(A)}$$

If the point is an outlier, its LRD will be much lower than the LRD of its neighbours, so its LOF will be larger

Compare the LRD of a point to the average LRD of its neighbours

LOF ~= 1 (Normal)
LOF >> 1 (Anomaly)

Contamination helps to set this threshold



Local Outlier Factor

$$LOF_2(A) = \frac{LRD_2(B) + LRD_2(C)}{\|N_2(A)\|} \times \frac{1}{LRD_2(A)} = \frac{0.5 + 0.667}{2} \times \frac{1}{0.667} = 0.87$$

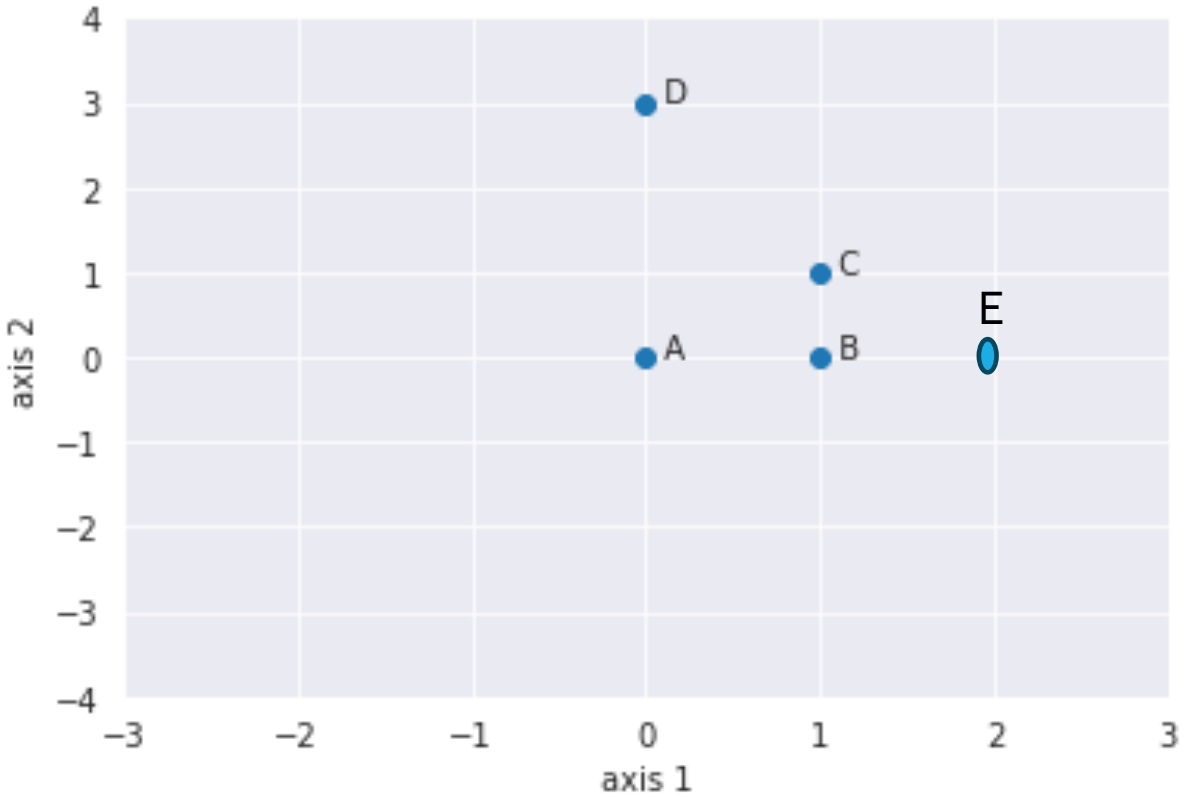
$$\mathrm{LOF}_2(B) = \frac{LRD_2(A) + LRD_2(C)}{\|N_2(B)\|} \times \frac{1}{LRD_2(B)} = \frac{0.667 + 0.667}{2} \times \frac{1}{0.5} = 1.334$$

$$\mathrm{LOF}_2(C) = \frac{LRD_2(B) + LRD_2(A)}{\|N_2(C)\|} \times \frac{1}{LRD_2(C)} = \frac{0.5 + 0.667}{2} \times \frac{1}{0.667} = 0.87$$

$$\mathrm{LOF}_2(D) = \frac{LRD_2(A) + LRD_2(C)}{\|N_2(D)\|} \times \frac{1}{LRD_2(D)} = \frac{0.667 + 0.667}{2} \times \frac{1}{0.337} = 2$$

Set of measured / derived values

- **Distance(A,Xi)**
  - Distance measured from a point A to any other point Xi
- **K = Closeness parameter**
  - Not measured, but set (e.g. 2, 10, 20)
- **K-Distance** – Distance to the Kth closest neighbour
- **N_K(A)** – Neighborhood of A (represented as a set), provided K
- **RD_K(A,Xj) – Reachability Distance**
  - For all Xj part of N_K(A), their individual capability of reaching A as Max(Distance(Xj,A) and K-Distance(Xj))
- **LRD_v(A) – Local Reachability Density of A** (to what extent its neighbors can reach it)
  - 1 / Average of all RD_v(A,Xj) in the neighborhood
- **LOF_v(A) – Local Outlier Factor of A** (to what extent its local reachability density compares to the local reachability density of its neighbors
  - LRD_v(A) / Average of all LRD_v(Xj), where Xj is part of N_v(A



I encourage you to practice

Source

Advantages

1.Effective in identifying outliers in datasets with varying densities of clusters.
2.Doesn’t require assumptions about the underlying distribution of the data.
3.Provides anomaly scores that can be used to rank the outliers.

Disadvantages

1.Sensitivity to the choice of parameters such as the number of neighbors (k_neighbors).
2.Can be computationally expensive for large datasets.
3.May require careful interpretation and adjustment of the anomaly scores threshold for outlier detection.

Four approaches

1) Interquartile Range
2) Z-score
3) K-NN
4) Local Outlier Factor

MANY more methods are developed…

DATA CLEANING

- Data wrangling
- Data quality
- Data cleaning
  - Duplicates
  - Invalid data
  - Outliers

Part 1

Part 2

I'll present missing data and data bias later (week 6)