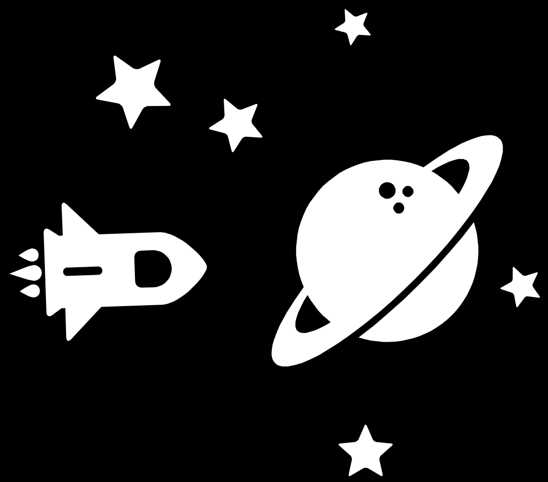
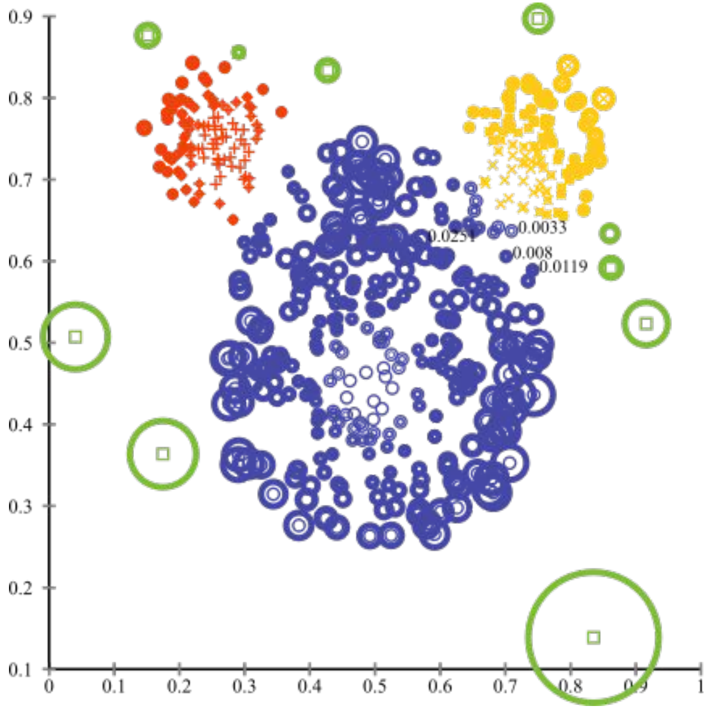
Outliers: Analysis & Detection

Presentation by Berk Sudan

Examples

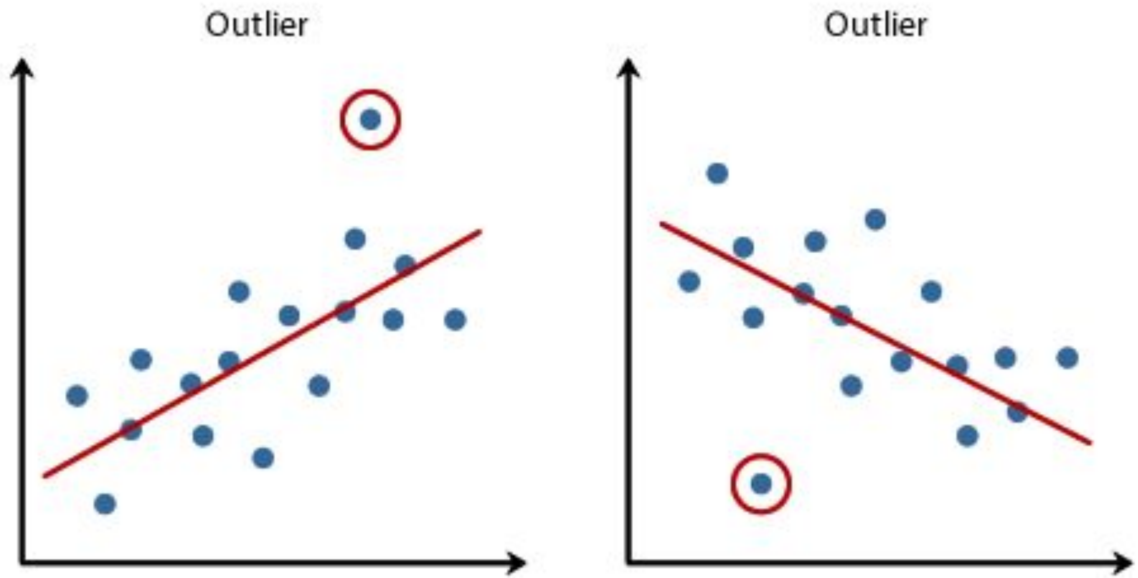


Clustering



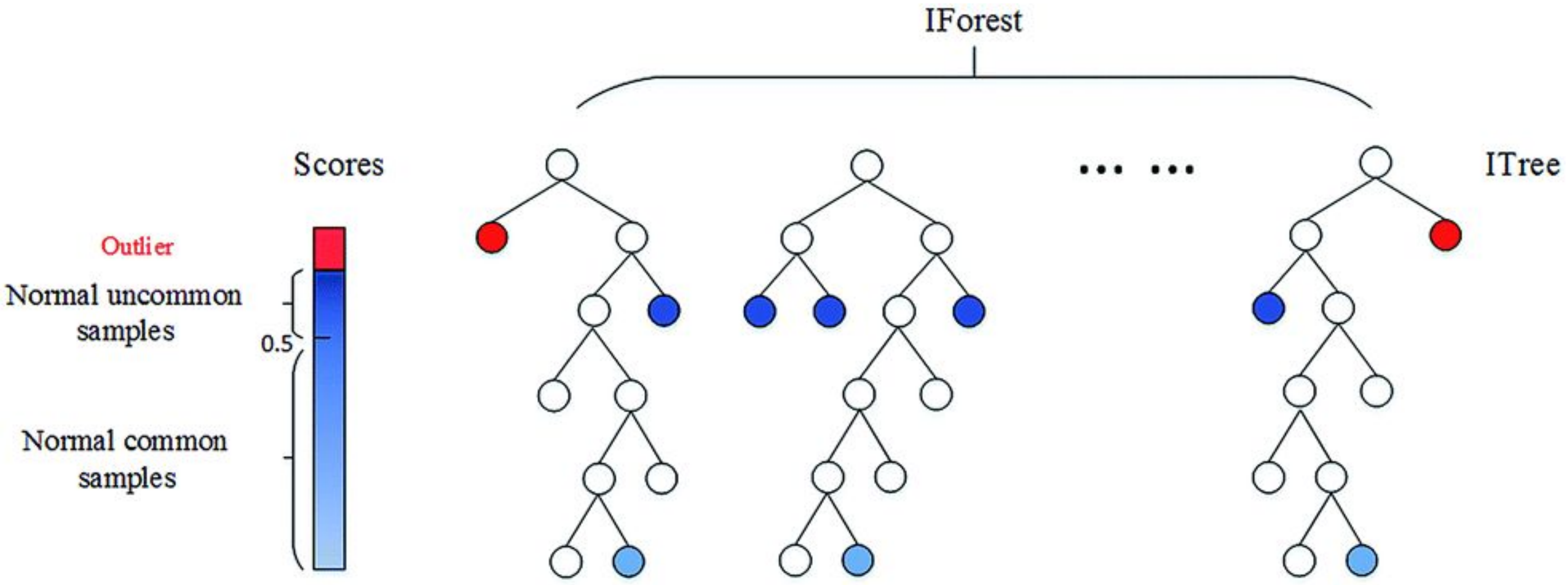
Ref: *https://stats.stackexchange.com/questions/160260/anomaly-detection-based-on-clustering*

Regression



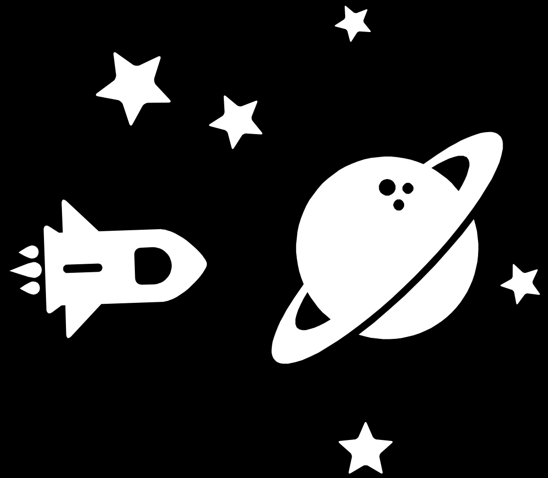
Ref: *https://datanee.com/2016/08/11/outlier-detection-an-overview-and-applications/*

Trees



Ref: *https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561*

What can outliers possibly be?



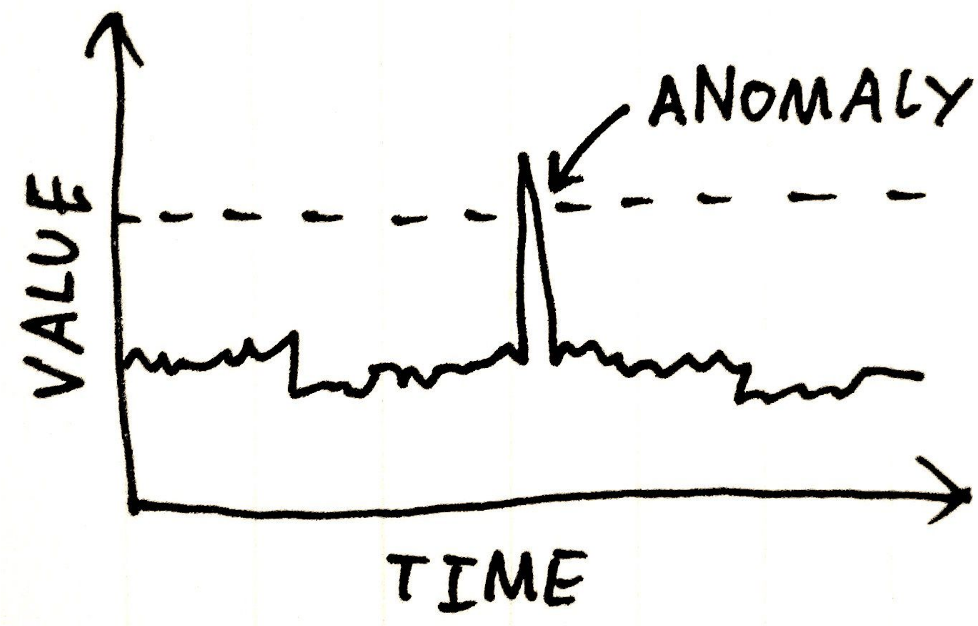
Possibility of Invalid Data - 1

Name	Gender	Year of Birth
Andrew NG	M	1976
Sarah Tan	F	2976

Possibility of Invalid Data - 2

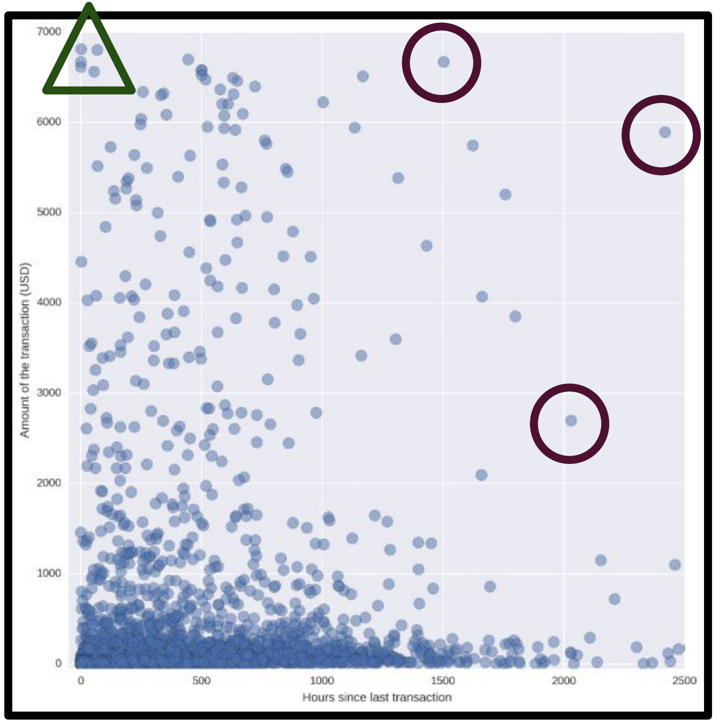
Gender	isPregnant	Age	isSmoking?
Male	Yes	25	Yes
Female	Yes	25	Yes
Male	No	25	No

Anomaly



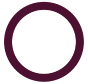
Ref: *https://medium.com/the-data-dynasty/anomaly-detection-in-google-analytics-a-new-kind-of-alerting-9c31c13e5237*

Fraud



X axis: Hours since last transaction

Y axis: Amount of the transaction (in USD)


: Outliers but not fraud


: Outliers & fraud

Ref: *https://blog.easysol.net/advanced-outlier-detection/*

What else?

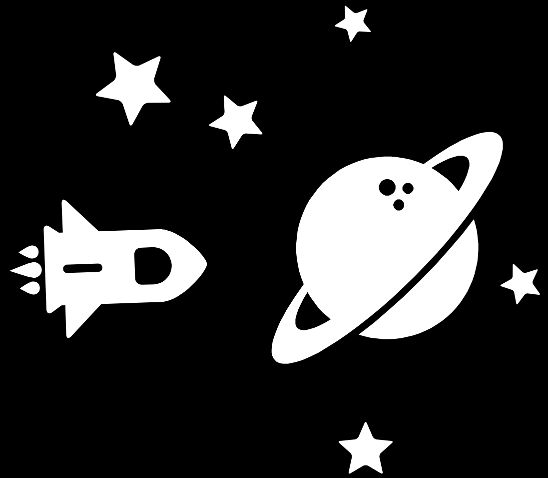
-> **Big variance, shifted mean** due to outliers

-> **Noise** in data

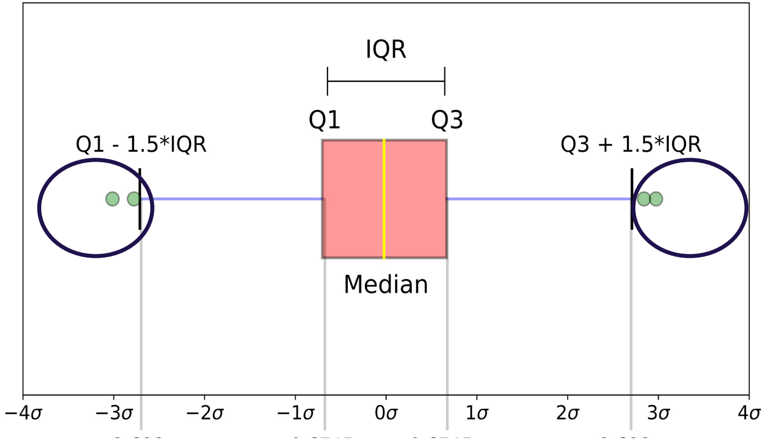
-> **Highly skewed** data distribution

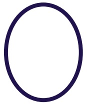
-> **Additional problems** in Distance-Based Algorithms (e.g, K-Means, K-NN, ...)

Methods: Uni-variate



Using Box-Plot





: outliers

For value X=x:

(x > (Q3 + **1.5*IQR**)) ∨ (x < (Q1 - **1.5*IQR**)) → (x is an **outlier**)

(x > (Q3 + **3*IQR**)) ∨ (x < (Q1 - **3*IQR**))→ (x is an **extreme-value**)

Ref: *https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51*

Using Standard Deviation

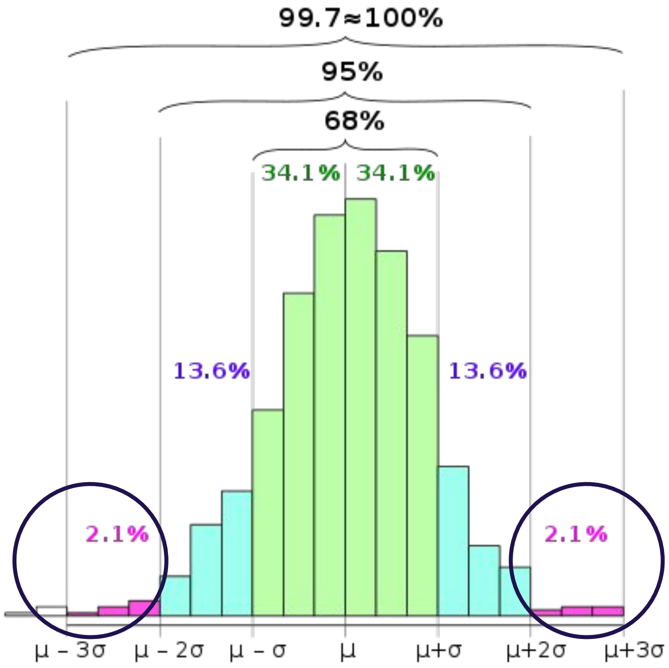
For value X=x:

(x > (μ + **2σ**)) ∨ (x < (μ - **2σ**)) →

(x is an **outlier**)

(x > (μ + **3σ**)) ∨ (x < (μ - **3σ**)) →

(x is an **extreme-value**)



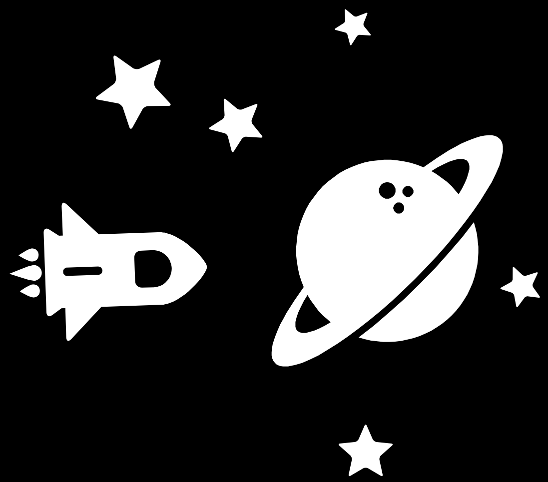
: outliers

Ref: *https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule*

Hard-Edges Method:

Data yielding outside of the **(1*th* - 99*th*) quantile/percentile interval** will be evaluated **as *outlier***.

Methods: Multi-variate



Using Distance based Approach

Euclidean Distance:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

FOR any point:

IF: distance (d) **>=** threshold_distance **AND**

fraction for non-neighbour sample (f) **>=** threshold_fraction

THEN: outlier!!

Using Distance based Approach - Example

$P_1 = (2,4)$
$P_2 = (3,2)$
$P_3 = (1,1)$
$P_4 = (4,3)$
$P_5 = (1,6)$
$P_6 = (5,3)$
$P_7 = (4,2)$

Find the outliers using the distance-based technique.

- The threshold distance is 3 (i.e, d >= 3)
- The threshold fraction f for non-neighbour sample is 4/7 (i.e, f >= 4/6)

Using Distance-based Approach - Example

Distance Matrix for Euclidean Distance

$$D_{7x7} = \begin{array}{c} \begin{array}{ccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 \end{array} \\ \begin{bmatrix} 0 & \sqrt{5} & \sqrt{10} & \sqrt{5} & \sqrt{5} & \sqrt{10} & \sqrt{8} \\ . & 0 & \sqrt{5} & \sqrt{2} & \sqrt{20} & \sqrt{5} & 1 \\ . & . & 0 & \sqrt{13} & 5 & \sqrt{20} & \sqrt{10} \\ . & . & . & 0 & \sqrt{18} & 1 & 1 \\ . & . & . & . & 0 & 5 & 5 \\ . & . & . & . & . & 0 & \sqrt{2} \\ . & . & . & . & . & . & 0 \end{bmatrix} \end{array} \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_6 \end{array}$$

Using Distance-based Approach - Example

- The threshold distance is 3 (i.e, d >= 3)
- The threshold fraction f is 4/6 (i.e, f >= 4/6)

Distance Matrix for Euclidean Distance

$$
D_{7x7} = \begin{array}{cc}
\begin{matrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 \end{matrix} & \\
\begin{bmatrix}
0 & \sqrt{5} & \sqrt{10} & \sqrt{5} & \sqrt{5} & \sqrt{10} & \sqrt{8} \\
. & 0 & \sqrt{5} & \sqrt{2} & \sqrt{20} & \sqrt{5} & 1 \\
. & . & 0 & \sqrt{13} & 5 & \sqrt{20} & \sqrt{10} \\
. & . & . & 0 & \sqrt{18} & 1 & 1 \\
. & . & . & . & 0 & 5 & 5 \\
. & . & . & . & . & 0 & \sqrt{2} \\
. & . & . & . & . & . & 0
\end{bmatrix} &
\begin{matrix}
p_1 \text{ -> f = 2} \\
p_2 \text{ -> f = 1} \\
p_3 \text{ -> f = 5} \\
p_4 \text{ -> f = 2} \\
p_5 \text{ -> f = 5} \\
p_6 \text{ -> f = 3} \\
p_6 \text{ -> f = 2}
\end{matrix}
\end{array}
$$

Using Distance-based Approach - Example

Distance Matrix for Euclidean Distance

- The threshold distance is 3 (i.e, d >= 3)
- The threshold fraction f is 4/6 (i.e, f >= 4/6)

$$
\begin{array}{rcl}
 & \begin{array}{ccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 \end{array} & \\
D_{7x7} = & \left[\begin{array}{ccccccc}
0 & \sqrt{5} & \sqrt{10} & \sqrt{5} & \sqrt{5} & \sqrt{10} & \sqrt{8} \\
. & 0 & \sqrt{5} & \sqrt{2} & \sqrt{20} & \sqrt{5} & 1 \\
. & . & 0 & \sqrt{13} & 5 & \sqrt{20} & \sqrt{10} \\
. & . & . & 0 & \sqrt{18} & 1 & 1 \\
. & . & . & . & 0 & 5 & 5 \\
. & . & . & . & . & 0 & \sqrt{2} \\
. & . & . & . & . & . & 0
\end{array}\right] &
\begin{array}{l}
p_1 \text{ -> f = 2} \\
p_2 \text{ -> f = 1} \\
p_3 \text{ -> f = 5} \\
p_4 \text{ -> f = 2} \\
p_5 \text{ -> f = 5} \\
p_6 \text{ -> f = 3} \\
p_6 \text{ -> f = 2}
\end{array}
\end{array}
$$

So, the outliers are: $\{P_3, P_5\}$

End of Presentation

Presented by Berk Sudan