Data Cleansing — 3

Faculty of Information Technology
Monash University, Australia

FIT5196 Week 8

Outline

Missing Data Mechanisms

- Describe relationships between measured variables and the probability of missing data
- Deciding upon the method for analysing missing values requires understanding about both the reasons for the missing values and the nature of the data for the missing observations.
- Three different missingness mechanisms:
  - Missing at random
  - Missing completely at random
  - Missing not at random

MAR, MCAR v.s. MNAR?

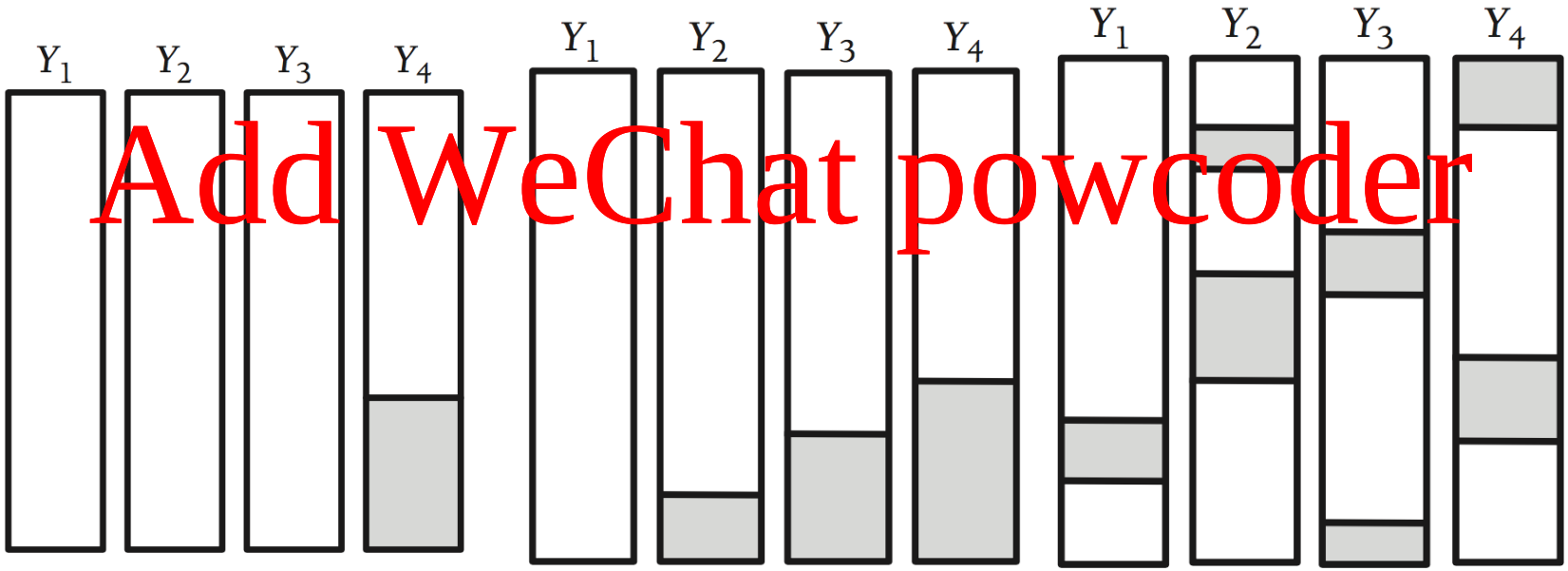
	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	—	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Example adopted from "Applied Missing Data Analysis" by Craig K. Enders.

Missing data Patten

A **missing data pattern** refers to the configuration of observed and missing values in a data set.

- The **univariate pattern** has missing values isolated to a single variable.
- A **monotone missing data** pattern is typically associated with a longitudinal study where participants drop out and never return.
- a **general pattern** has missing values dispersed throughout the data matrix in a haphazard fashion.



Methods for handling missing values

- Deletion methods
  - Listwise deletion
  - Pairwise deletion
- Imputation methods
  - Mean imputation
  - Regression imputation
  - Stochastic regression imputation
  - Hot-deck imputation
  - Last observation carrie forward

Outline

Outliers: the definition

- What is an outlier?
  - Definition of Hawkins: "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, D. 1980. Identification of Outliers. Chapman and Hall.)
  - Definition of Pearson: "An outlier is a data point that appears to be inconsistent with the nominal behavior exhibited by most of the other data points in a specified collection."

Outliers: the definition

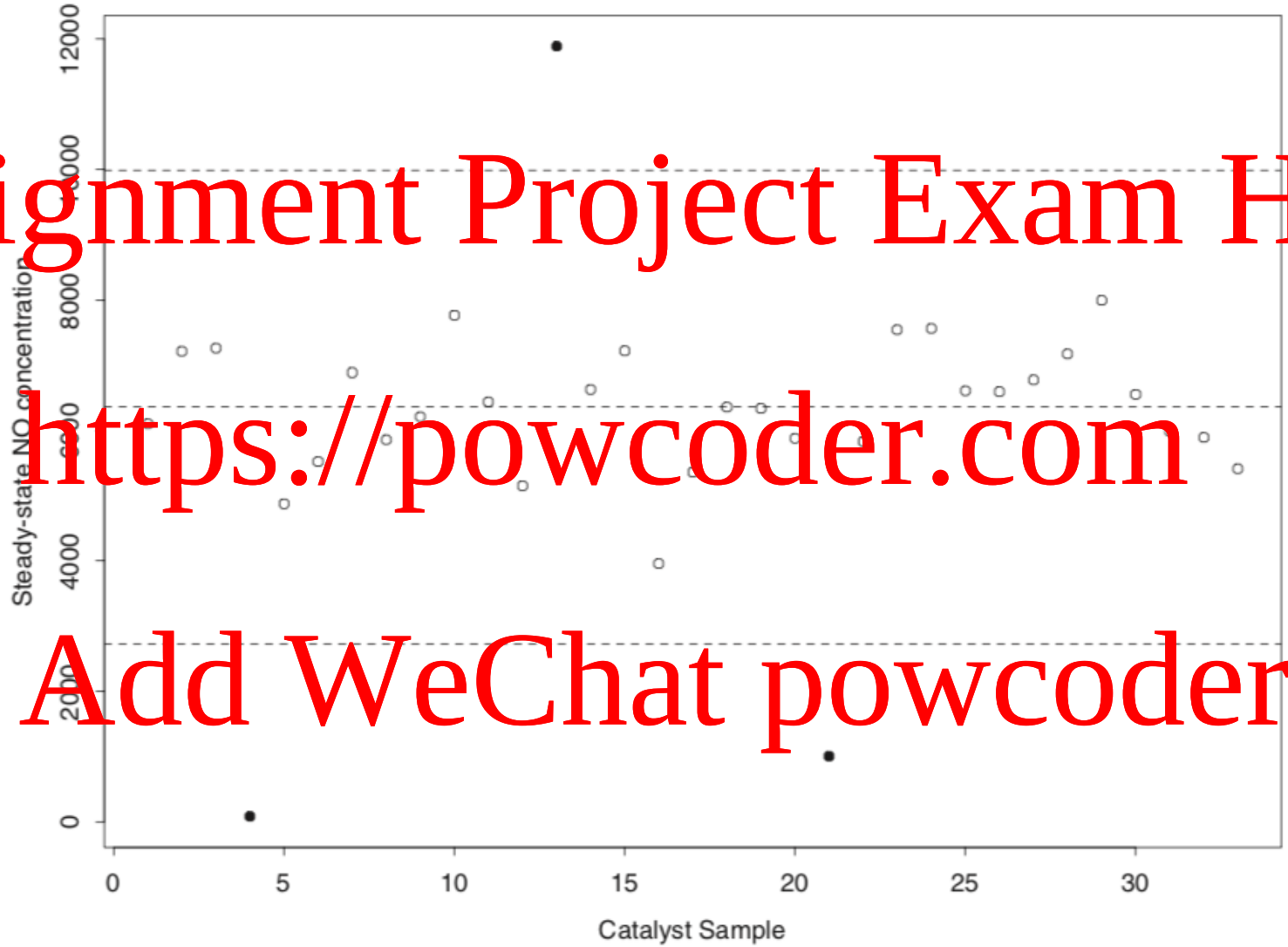


Figure is from Chapter 2 of "Mining Imperfect Data". Outliers detected with the Hampel identifier are catalyst samples 4, 13, and 21 and are marked with solid circles, The median value and upper and lower Hampel identifier detection limits are shown as dashed lines.

Outliers

- An outlier often contains useful information about abnormal characteristics of the systems and entities that impact the data generation process.
  - Intrusion detection systems: Unusual behaviour shown in the operating system calls, network traffic, or other user action.
  - Credit-card fraud: Unauthorized use of a credit card may show different patterns, such as buying sprees from particular locations or very large transactions.
  - Medical Analysis: Unusual patterns in MRI, PET and ECT data typically reflect disease conditions
  - Law enforcement: Determining fraud in financial transactions, trading activity, or insurance claims typically requires the identification of unusual patterns in the data generated by the actions of the criminal entity.

Outliers: the impact

- Outliers can increase the error variance and reduces the power of statistical tests.
- If the outliers are non-randomly distributed, they can decrease normality.
- Outliers can bias or influence estimates that may be of substantive interest
- Outliers can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

Outliers: the impact

Example:

8,7,9,9,6,5,8,9,8,8,9	8,7,9,9,6,5,8,9,8,8,9,100
mean = 7.8	mean = 15.5
median = 8	median = 8
mod = 8	mod = 8
sd = 1.328	sd = 26.641

Types of outliers

- Univariate outlier: concerns the distribution of a single variable
- Multivariate outlier: concerns outliers in an n-dimensional space.

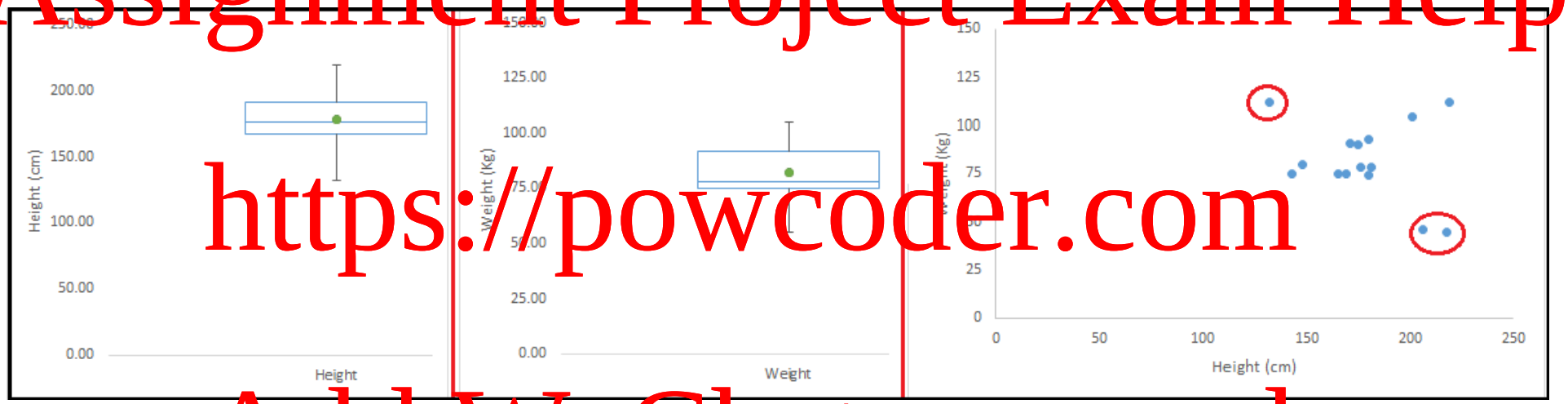


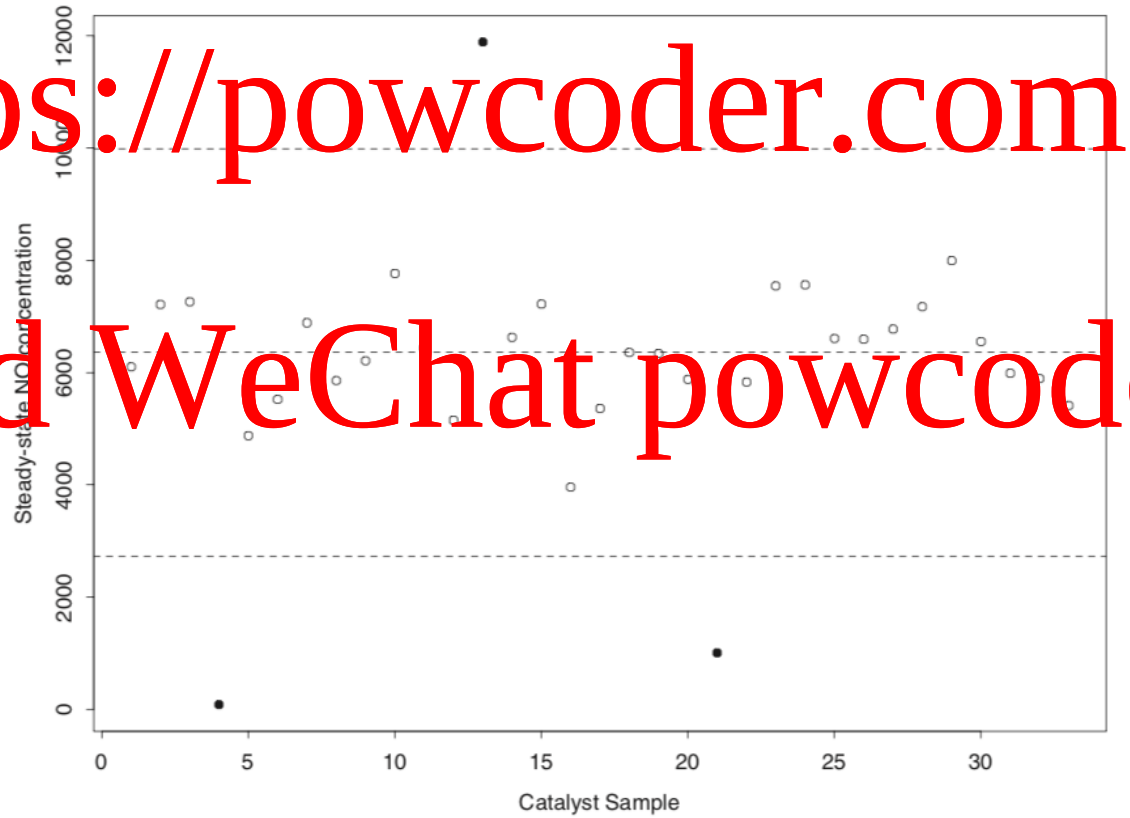
Figure is from "A Comprehensive Guide to Data Exploration"

Univariate Outliers

- Based on the notion that "most " of the data should exhibit approximately the same value c, the observed sequence of data $\{x_k\}$ can be modelled as

$$x_k = c + e_k$$

where $\{e_k\}$ is a sequence of deviations about the nominal value c.



Univariate Outliers

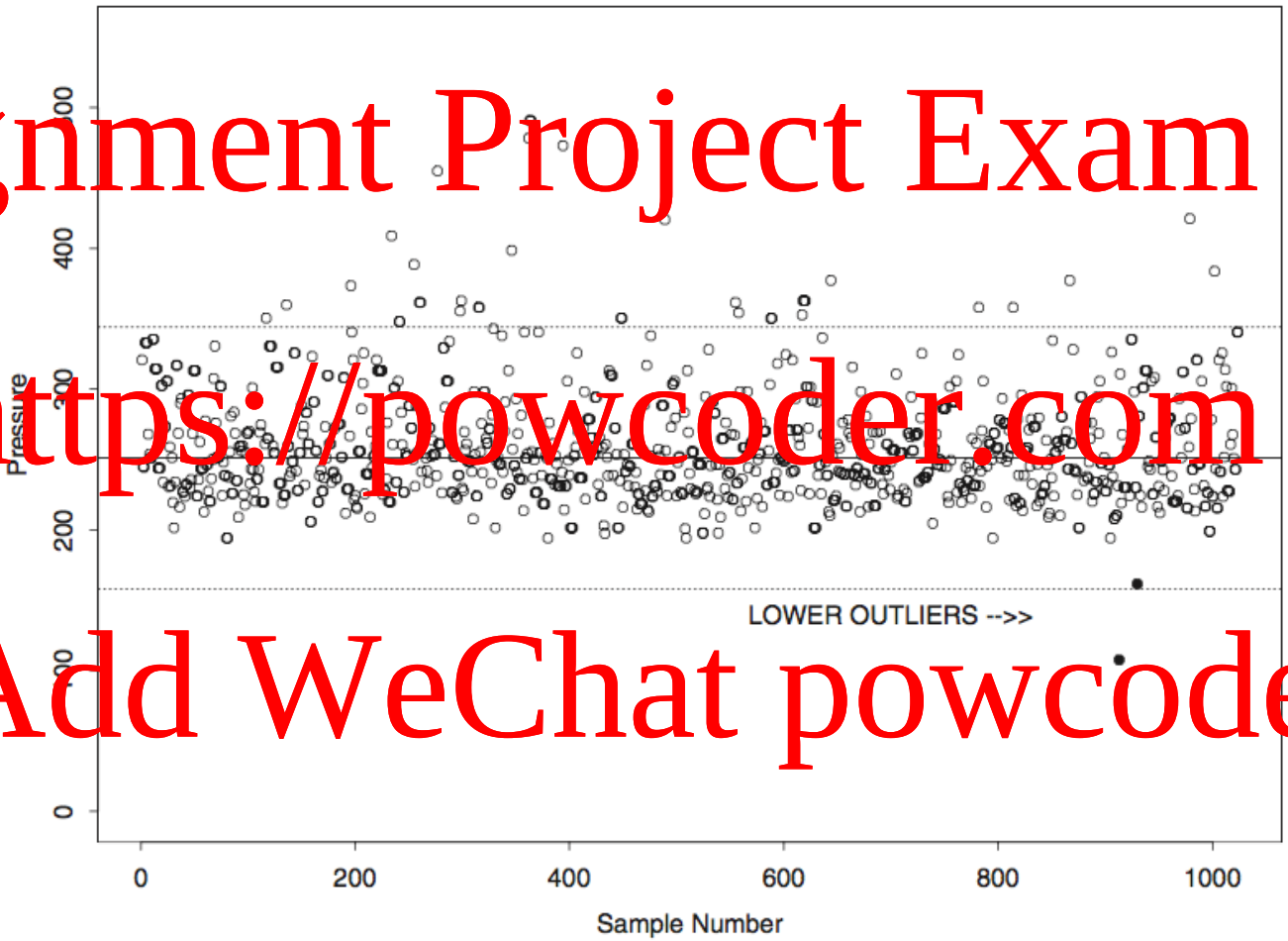


Figure is from Chapter 2 of "Mining Imperfect Data"

- Distinguish between lower outliers and upper outliers

Multivariate outliers

- Multivariate outlier: concerns outliers in an n-dimensional space.
  - A multivariate outlier in a sequence $\{x_k\}$ of vectors corresponds to a vector x_i whose individual components are significantly discordant with the intercomponent relations exhibited by the majority of the other data values.

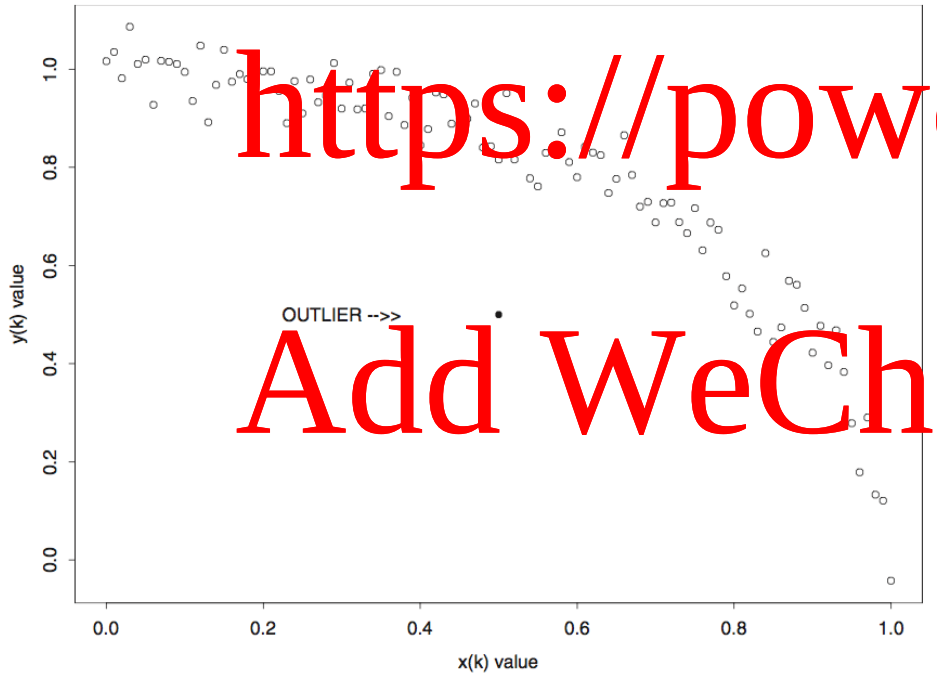


Figure is from Chapter 2 of "Mining Imperfect Data"

- the intercomponent relation: $y \approx \sqrt{1 - x^2}$
- $0 \lessapprox x, y, \lessapprox 1$

How to detect Univariate Outliers

- Problem formulation: given a sequence of observed data $\{x_k\}$, a reference value x_0, and a measure of variation ζ computed from $\{x_k\}$, detect outliers according to

$$|x_k - x_0| > t\zeta$$

where t is a threshold parameter.

- Questions
  - How do we define the nominal data reference value x_0?
  - How do we define the scale of natural variation ζ ?
  - How do we choose the threshold parameter t?

Three Outlier Detection Methods

- Choices for the nominal reference value x_0
  - mean: $\bar{x}$
  - median: $x^\dagger$
- Choices for the measure of variation ζ
  - the standard deviation: σ
  - The median absolute deviation(MAD) scale estimator S:

$$S = 1.4826 \times median\{|x_k - x^\dagger|\}$$

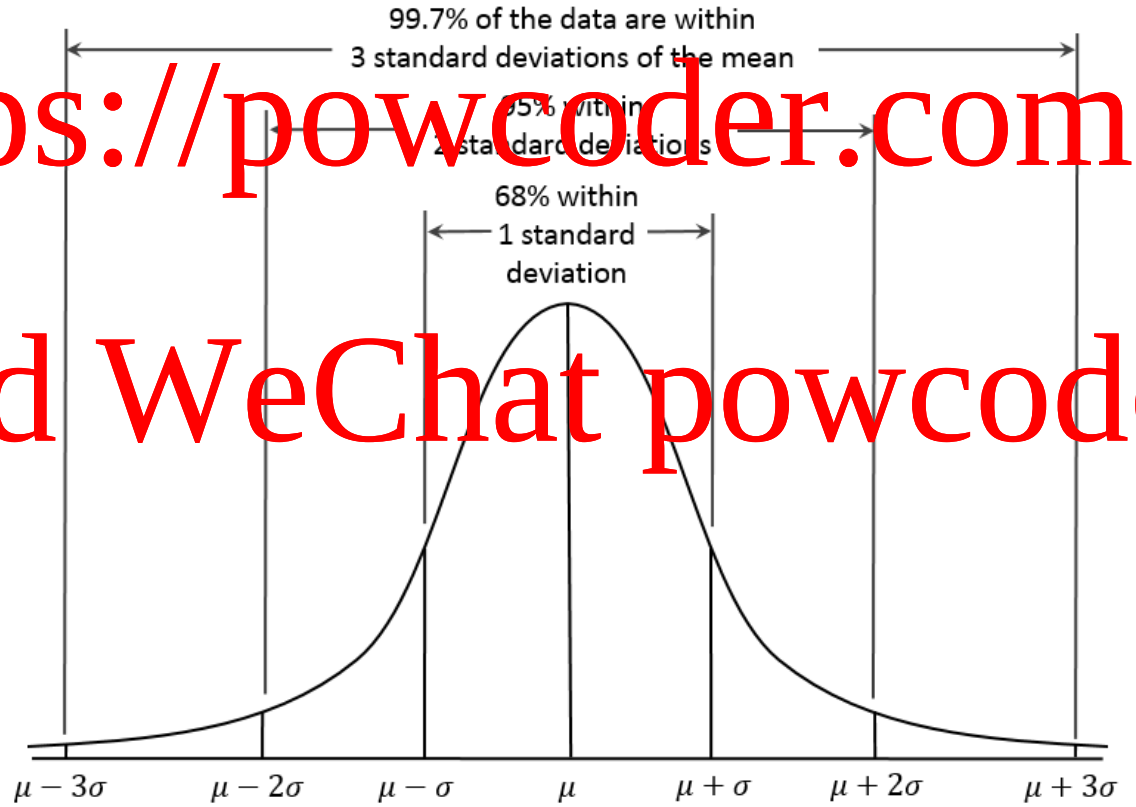
  - The Interquartile Range (IQR)

$$IQR = Q_3 - Q_1$$

- Combine the choices
  - The 3σ edit rule: $x_0 = \bar{x}$, $\zeta = \sigma$
  - The Hampel identifier: $x_0 = x^\dagger$, $\zeta = S$
  - The standard boxplot outlier rule: $x_0 = x^\dagger$, $\zeta = IQR$

The 3σ edit rule

- Basic idea: if a data sequence $\{x_k\}$ is well approximated by an i.i.d. sequence of Gaussian random variables with mean μ and standard deviation σ, the probability of observing a value x_k farther than three standard deviations from the mean is only about 0.3%.



The 3σ edit rule

- x_k is an outlier if

$$|x_k - \bar{x}| > 3\sigma$$

  Also known as the extreme studentized deviation (ESD) identifier (Davies and Gather, 1993)
- Problems?

The 3σ edit rule

- x_k is an outlier if

$$|x_k - \bar{x}| > 3\sigma$$

  Also known as the extreme studentized deviation (ESD) identifier (Davies and Gather, 1993)
- Problems?
  - The presence of outliers in the dataset can cause substantial errors in estimating
    - the mean
    - the standard deviation

8,7,9,9,6,5,8,9,8,8,9	8,7,9,9,6,5,8,9,8,8,9,100
mean = 7.8	mean = 15.5
avedev = 0.99	avedev = 14.08
sd = 1.328	sd = 26.641

The Hampel Identifier

- Basic idea:
  - $x_0 = x^\dagger$
  - $S = 1.4826 \times median\{|x_k - x^\dagger|\}$
  - x_k is an outlier if

$$|x_k - x^\dagger| > 3S$$

The Hampel Identifier

- Basic idea:
  - $x_0 = x^\dagger$
  - $S = 1.4826 \times median\{|x_k - x^\dagger|\}$
  - x_k is an outlier if

$$|x_k - x^\dagger| > 3S$$

- Why use median and MAD
  - lower outlier-sensitivities than mean and standard deviation

8,7,9,9,6,5,8,9,8,8,9	8,7,9,9,6,5,8,9,8,8,9,100
median = 8	median = 8
MAD = 1	MAD = 1

The Hampel Identifier

- Basic idea:
  - $x_0 = x^\dagger$
  - $S = 1.4826 \times median\{|x_k - x^\dagger|\}$
  - x_k is an outlier if

$$|x_k - x^\dagger| > 3S$$

- Drawbacks:
  - the MAD scale estimate is identically zero if more than 50% of the data observations x_k have the same value.

Quartile-based Detection and Boxplots

Q0	the minimum
Q1	bigger than 25% of the data points
Q2	the median
Q3	bigger than 75% of the data points
Q4	the maximum

- For a symmetric distribution,

$$
\begin{aligned}
IQR &= Q3 - Q1 \\
x^{\dagger} &= \frac{Q3 + Q1}{2} \\
Q3 &= x^{\dagger} + IQR/2 \\
Q1 &= x^{\dagger} - IQR/2
\end{aligned}
$$

- The observation suggests
  - $x_0 = x^{\dagger}$
  - $\zeta = IQR$

Quartile-based Detection and Boxplots

- Symmetric boxplot rule

$$|x_k - x^\dagger| \geq t \times IQR$$

- Asymmetric boxplot rule

$$x_k > Q3 + t \times IQR \quad \Rightarrow \quad x_k \text{ is an upper outlier}$$
$$x_k < Q1 - t \times IQR \quad \Rightarrow \quad x_k \text{ is an lower outlier}$$

Multivariate Outlier Detection

MONASH University

- Linear models
  - Residuals, i.e., the distances of the data points from this hyperplane, are used to quantify the outlier scores.
- Proximity-based models
  - Outliers are defined as those points that do not lie in the dense regions.
    - Clustering methods: segment the data points
    - Density based methods: segment the data space.

Linear models

- linear regression model

$$y = \sum_{i=1}^{d} w_i x_i + w_{d+1} + \epsilon_j$$

- Learning objective: minimise the error between the true value of the predicted value of y

$$\sum_j \epsilon_j^2 = \sum_j ((\sum_{i=1}^{d} w_i x_{j,i} + w_{d+1}) - y_j)^2 \quad (1)$$

$$= ||\mathbf{Dw} - \mathbf{y}||^2 \quad (2)$$

where $\mathbf{D}$ is $\mathbf{N} \times (d+1)$ data matrix, W is the coefficients, $\mathbf{y}$ is a vector N true response values.

- Closed form soluction

$$\mathbf{w} = (\mathbf{D}^t\mathbf{D} + \alpha\mathbf{I})^{-1}\mathbf{D}^t\mathbf{y}$$

Linear models

- Regression with and without outliers

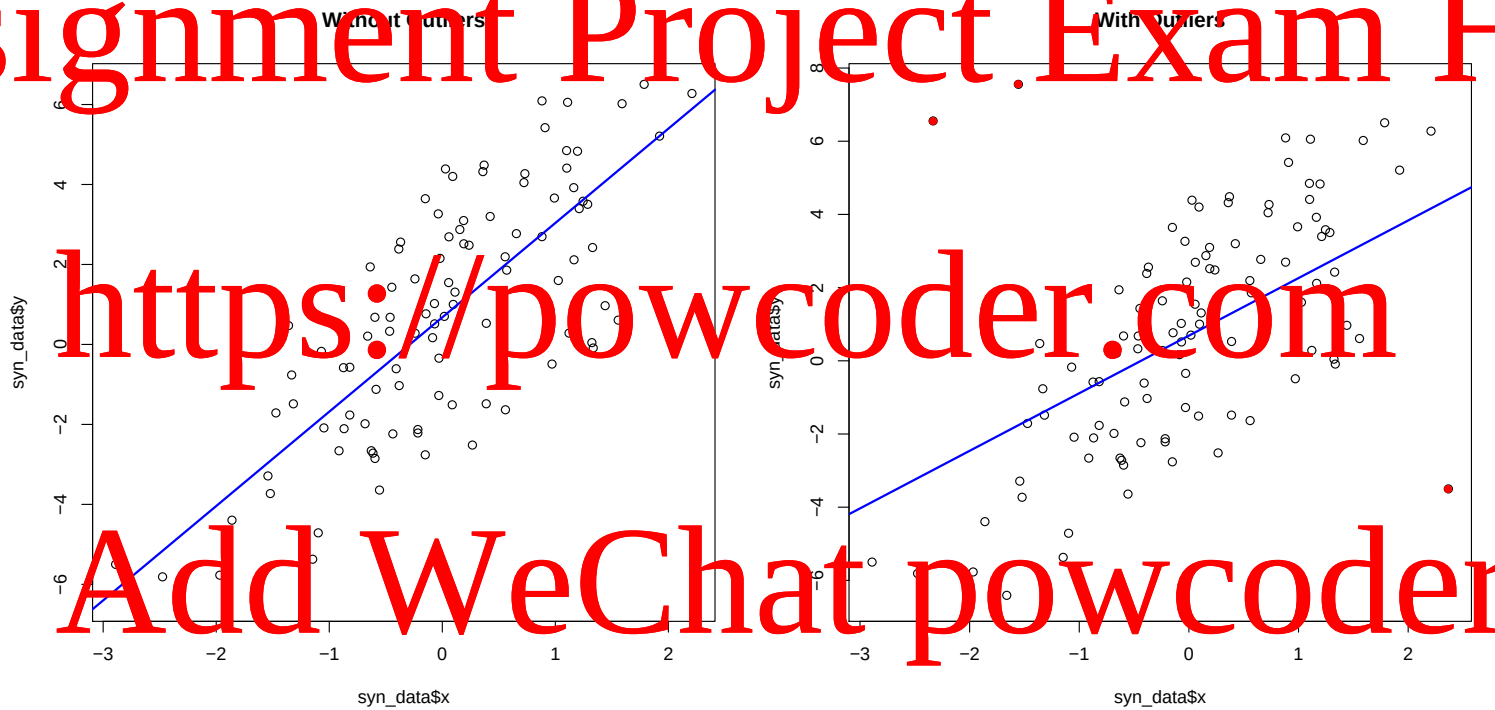
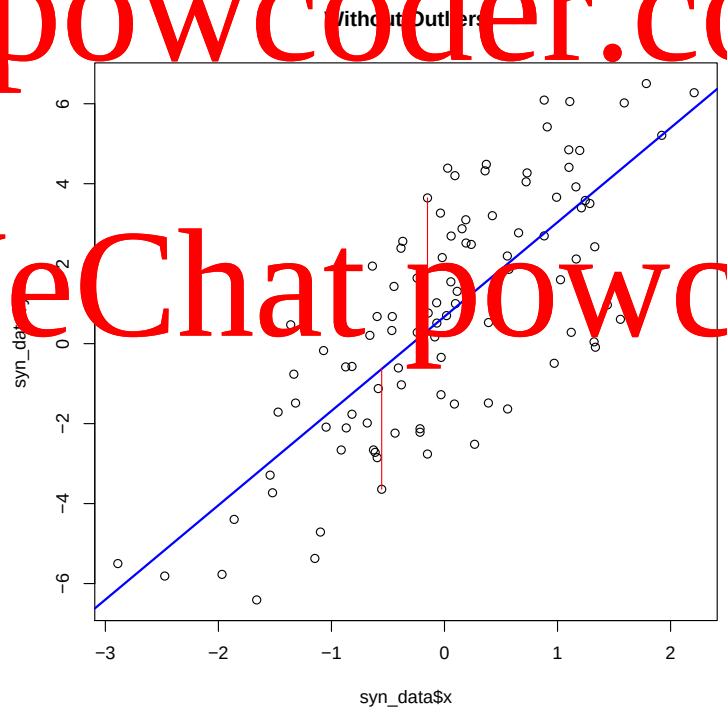


Figure: $\mathbf{y} = 2\mathbf{x} + 0.5 + \boldsymbol{\epsilon}$

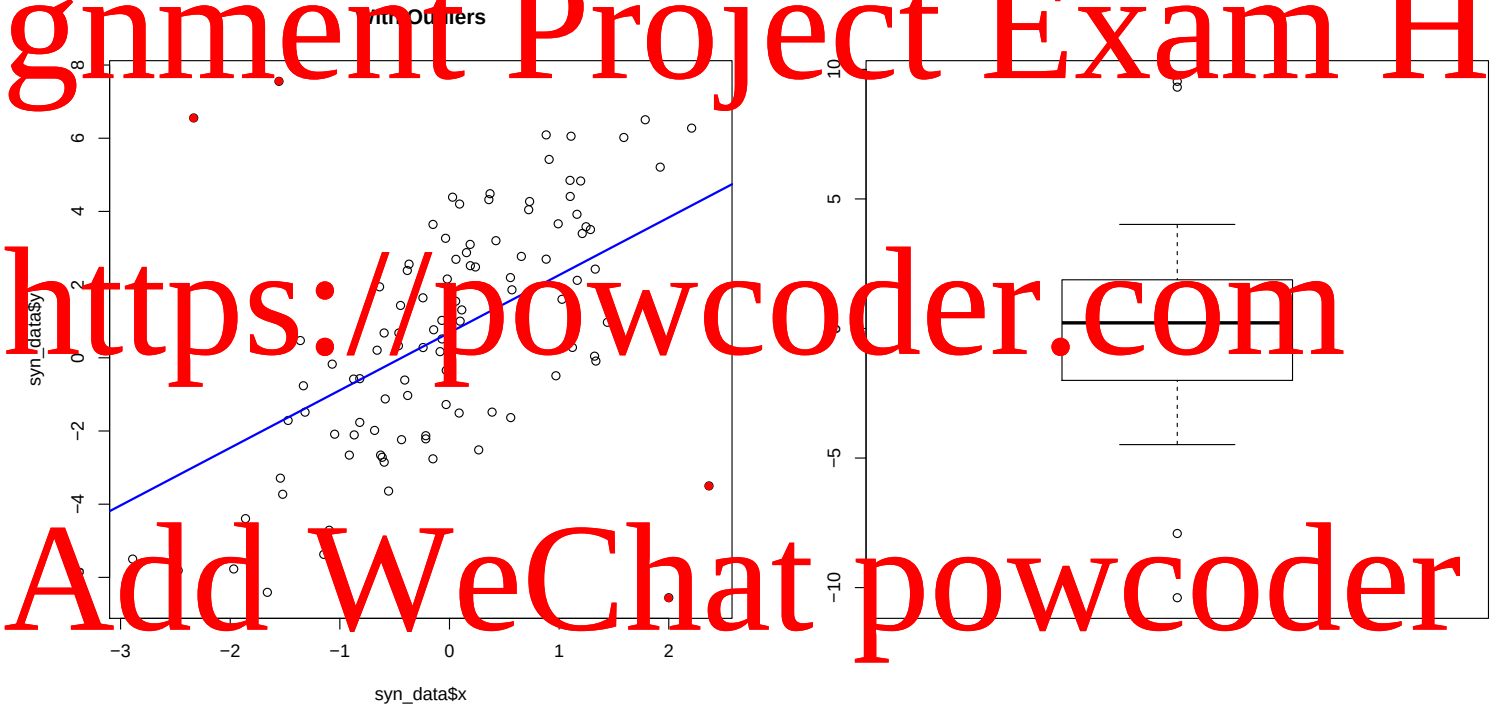
Linear models

- Outliers are, after all, values that deviate from expected (or predicted) values on the basis of a particular model
- Goal: find lower-dimensional subspaces, in which the outlier points behave very differently from other points
  - The residual ϵ_j provides useful information about the outlier score of the data point j.

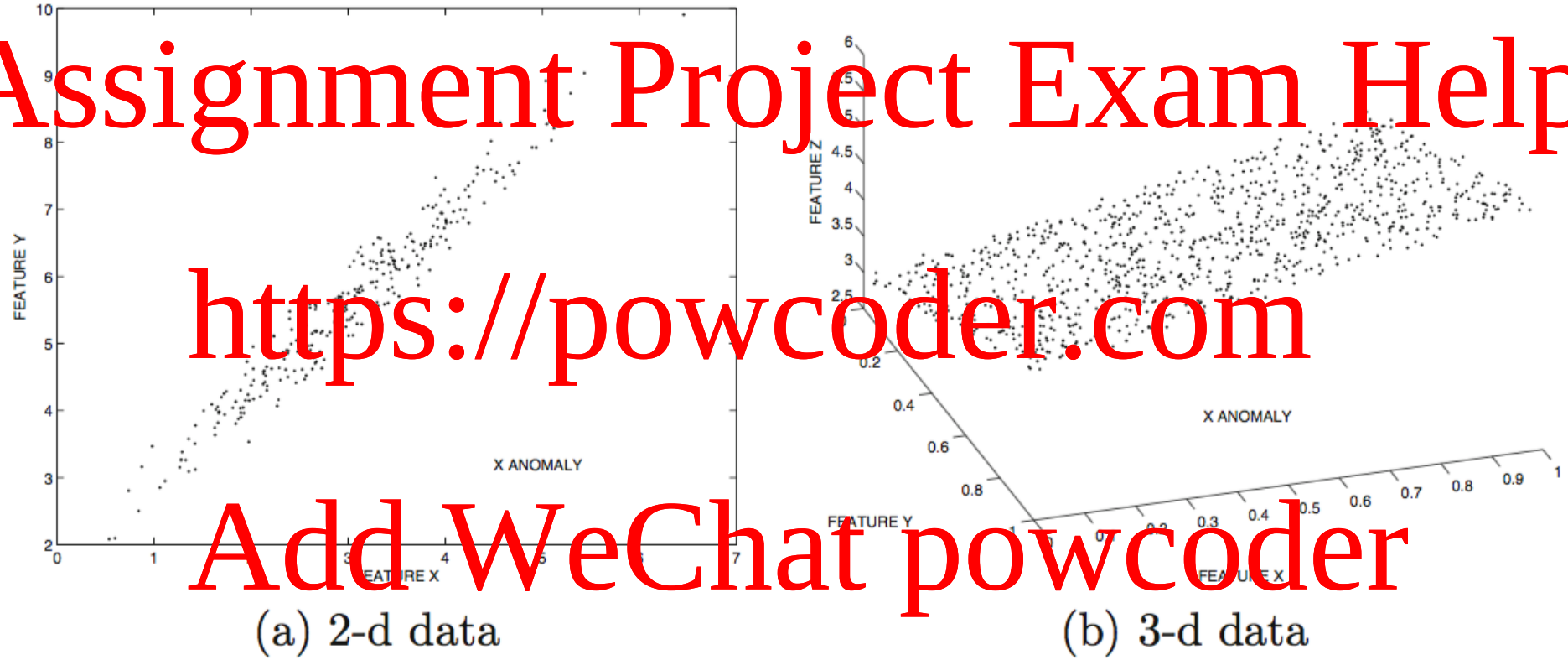


Linear models

- Using boxplot



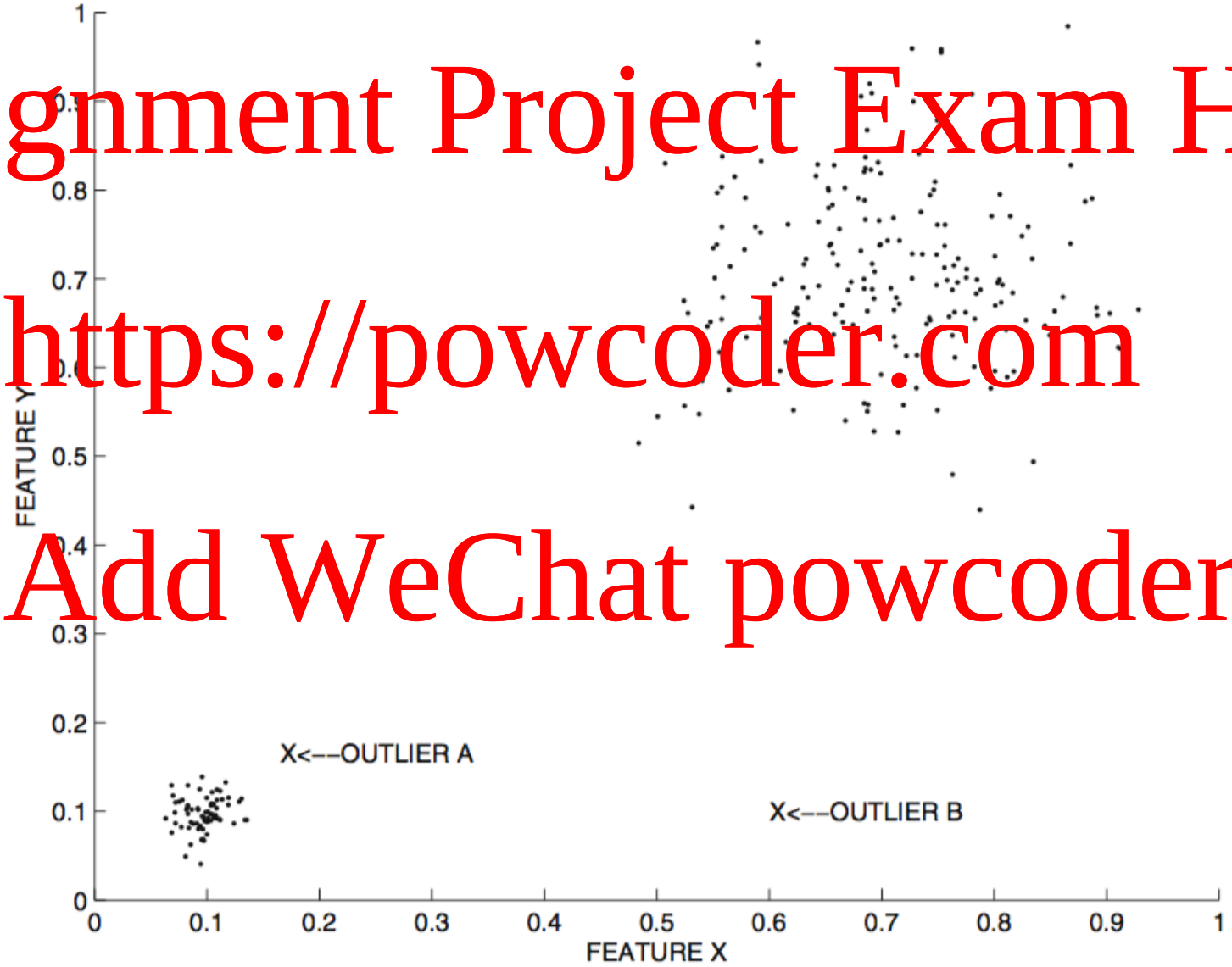
Density-Based Outliers



(a) 2-d data (b) 3-d data

Figure: Figure from "Outlier Analysis", second edition by Charu C. Aggarval

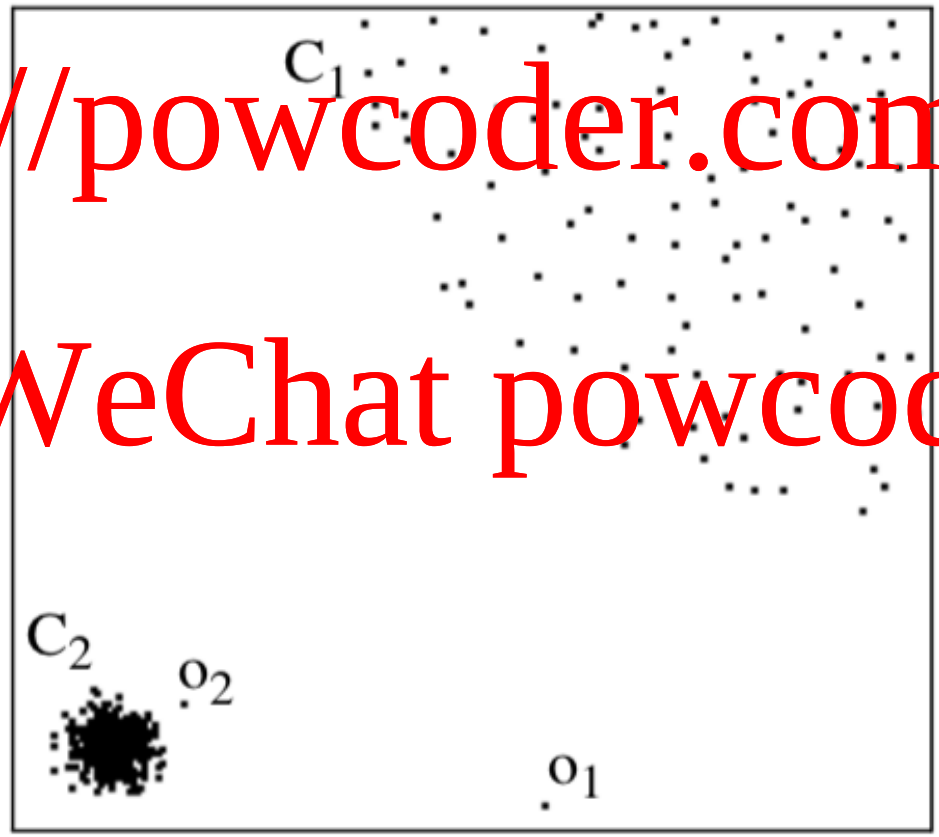
Density-Based Outliers



Density-Based Outliers

- Distance-based method:
  - An object p in a dataset D is a $DB(pct, dmin)$-outlier if at least percentage pct of the objects in D lies greater than distance $dmin$ from p,

$$|\{q \in D | d(p, q) \leq dmin\}| \leq (100 - pct) \times |D|.$$



Density-Based Outliers

MONASH University

- Local Outlier Factor (LOF)
  - k-distance of an object p, denoted as $d_k(p)$ is defined as the distance $d(p, o)$ between p and an object $o \in D$ such that:
    1. for at least k objects $o' \in D\backslash\{p\}$ it holds that $d(p, o') \leq d(p, o)$, and
    2. for at most $k-1$ objects $o \in D\backslash\{p\}$ it holds that $d(p, o') < d(p, o)$.
  - k-distance neighborhood of an object p

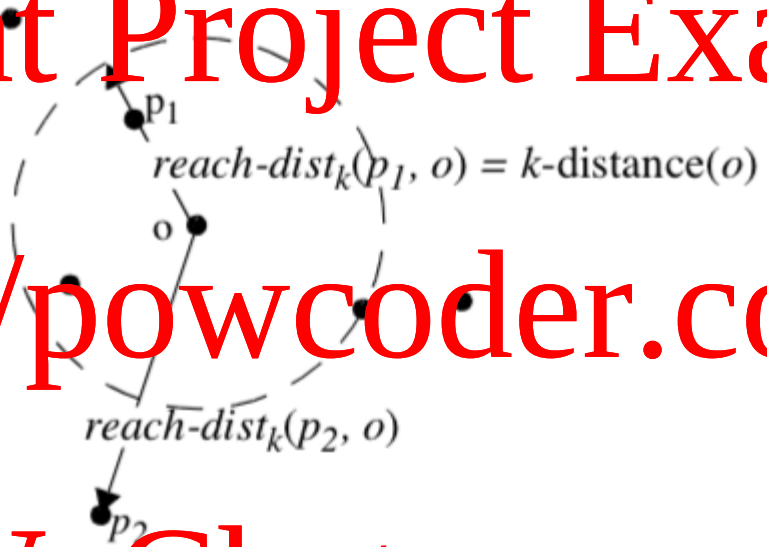
$$N_{d_k(p)}(p) = \{q \in D\backslash\{p\} \mid d(p, q) \leq d_k(p)\}$$

  - Reachability distance of an object p w.r.t. object o

$$d_{rk}(p, o) = \max(d_k(o), d(p, o))$$

Density-Based Outliers

- Local Outlier Factor (LOF)



where $reach - dist_k(p_1, o) = d_{r,k}(p_1, o)$ and $k - distance(o) = d_k(o)$

Density-Based Outliers

- Local Outlier Factor (LOF)
  - Local reachability density of an object p

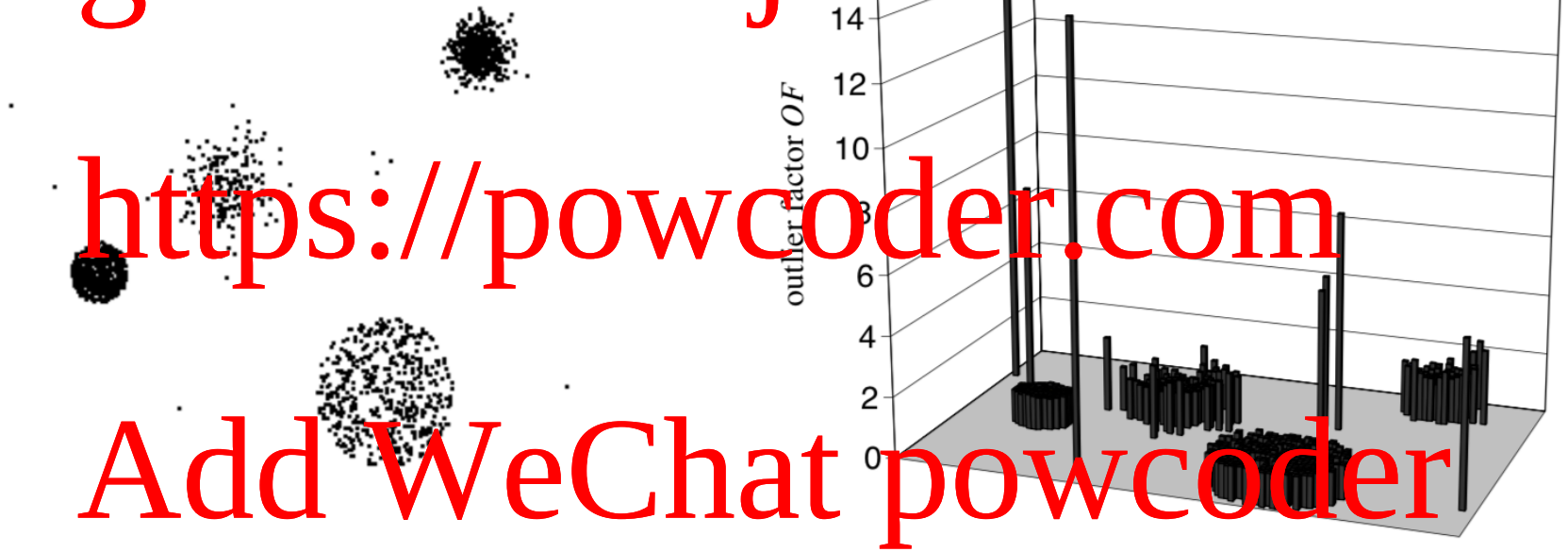
$$lrd_k(p) = \frac{1}{\left(\frac{\sum_{o \in N_k(p)} d_{r,k}(p,o)}{|N_k(p)|}\right)}$$

  - LOF of an object p

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{| N_k(p) |}$$

Density-Based Outliers

- Local Outlier Factor (LOF)



Compare different outlier detection methods

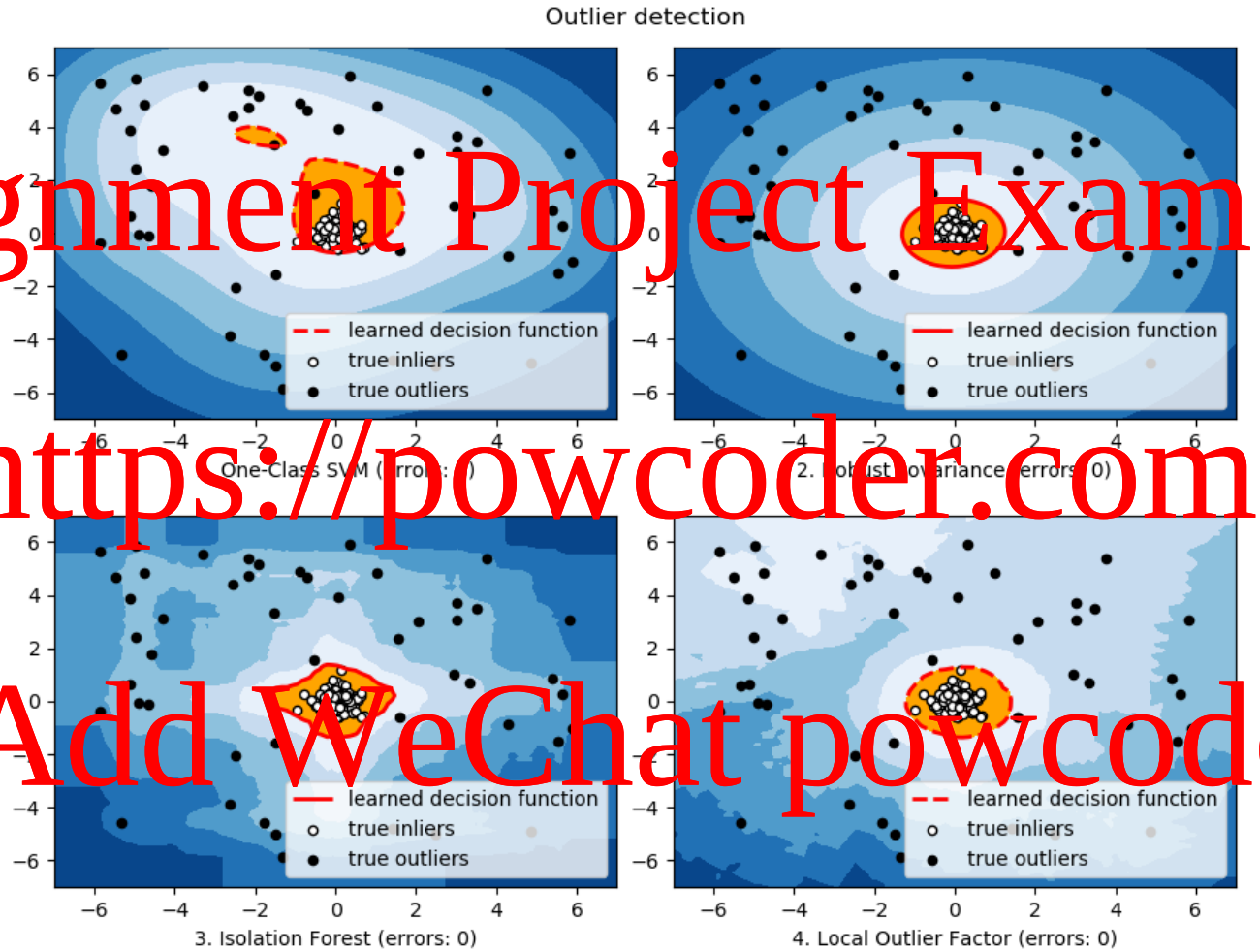


Figure: Figures from http://scikit-learn.org/

Compare different outlier detection methods

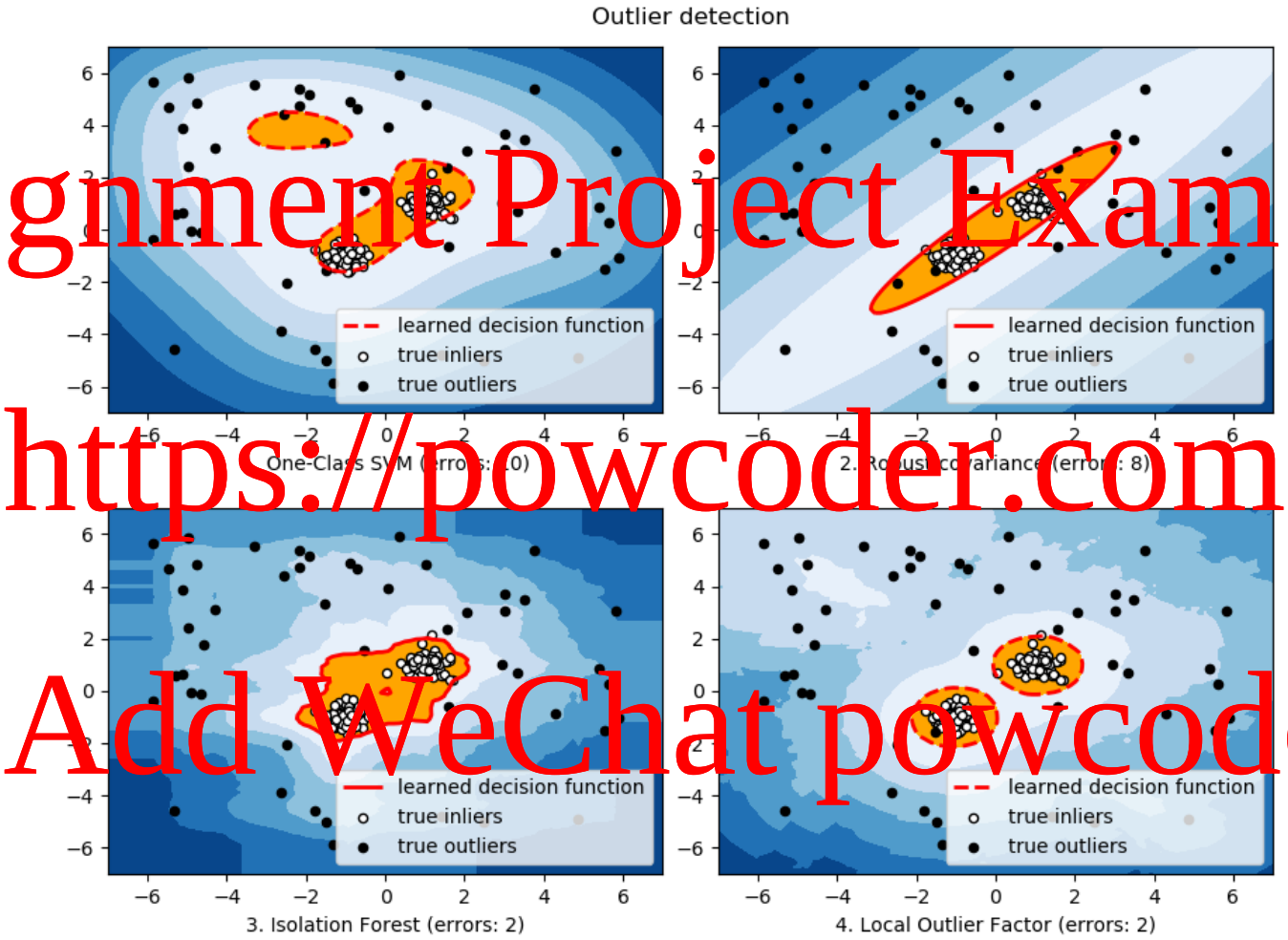


Figure: Figures from http://scikit-learn.org/

Compare different outlier detection methods

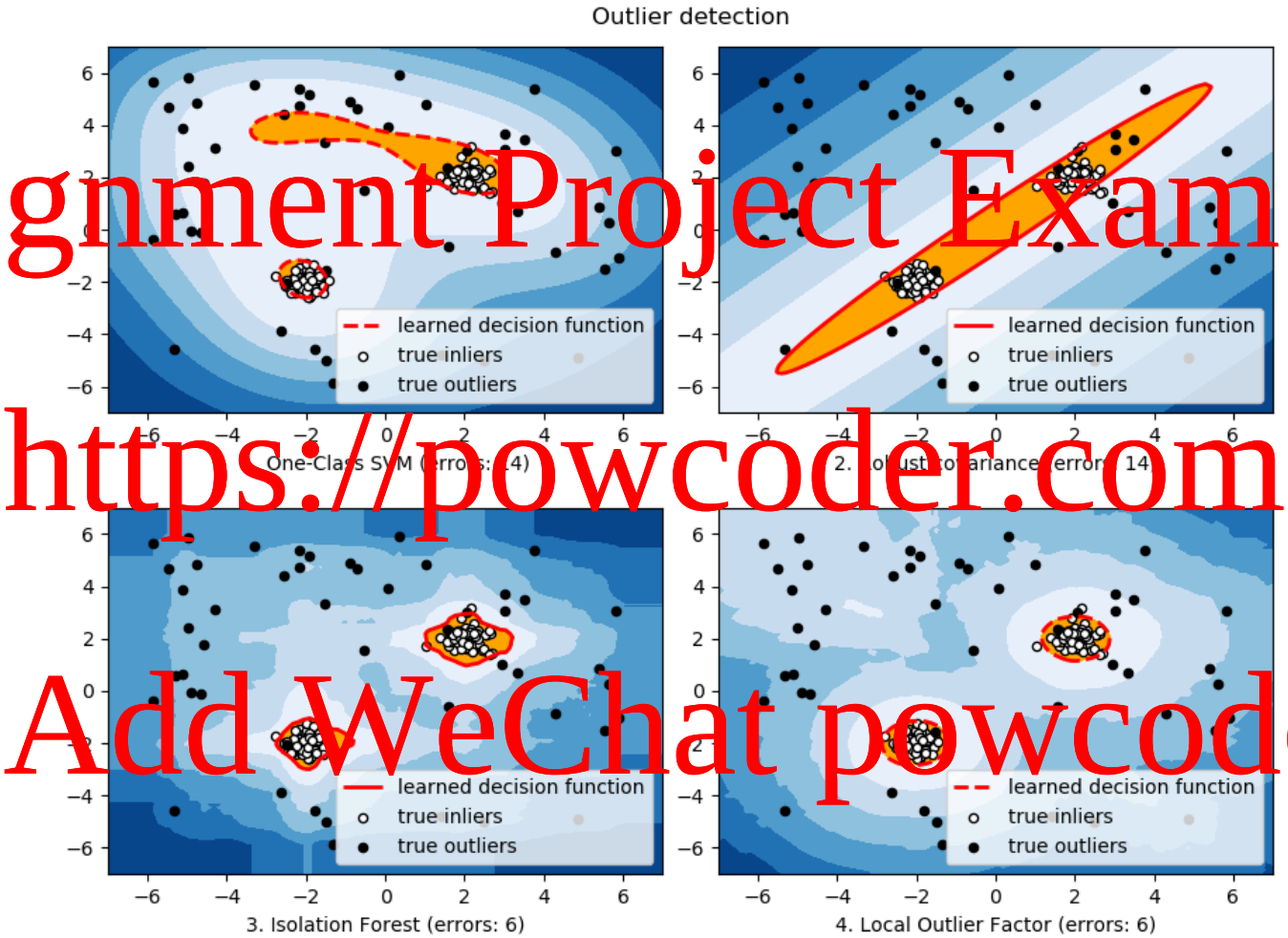


Figure: Figures from http://scikit-learn.org/

Summary

- Types of outliers
- Univariate outlier detection method
  - the 3σ edit rule
  - the Hampel identifier
  - the Quartile-based detection
- Multivariate outlier detection method
  - Linear model
  - Local Outlier factor