DATA CLEANING

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

- Data cleaning involves the following processes:

i) Manage missing data.

ii) Manage inconsistent data.

iii) Manage outliers.

- If users know the data is 'unclean/dirty', they will not trust the results that you present.

- 'Dirty' data can cause confusion/difficulty in data mining procedures, also resulting in unreliable results.

- Missing data is due to various factors, including:

i) Individual errors during data entry/recording

Example: Human Factor.

ii) There is a malfunction of the data recorder

Example: Meteorology sensor.

iii) Customer refuse to provide information.

Example: Survey sampling, Cencus.

iv) Such data does not exist naturally.

Example: Attribute for driving license number, some respondents do not have a driving license.

MANAGE MISSING DATA:

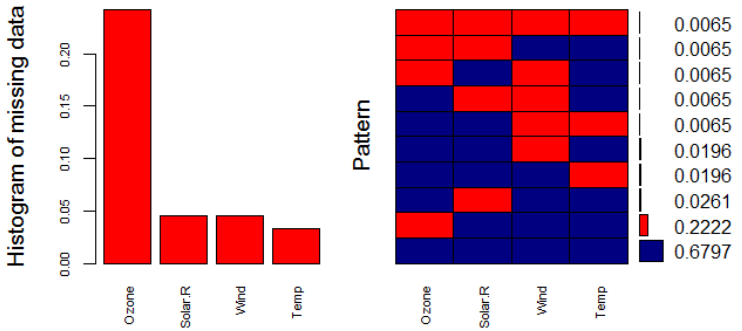
- Several approaches can be used in managing missing data, among them are:

1. Identify the patterns and trend of missing data:

- It aims to get some insight about the behaviors of missing data in the data set.

2. Remove observations that contain a missing data :

- If the amount of our data is large and the percentage of missing data is small, this technique could be appropriate.
- However, it is not effective if the missing data is quite large.
- Observations containing missing data may contain important information in other attributes. Missing data needs to be estimated.
- However, some of the missing data is not suitable to be estimated and it needs to be removed. This depends on the domain knowledge of the analyst.

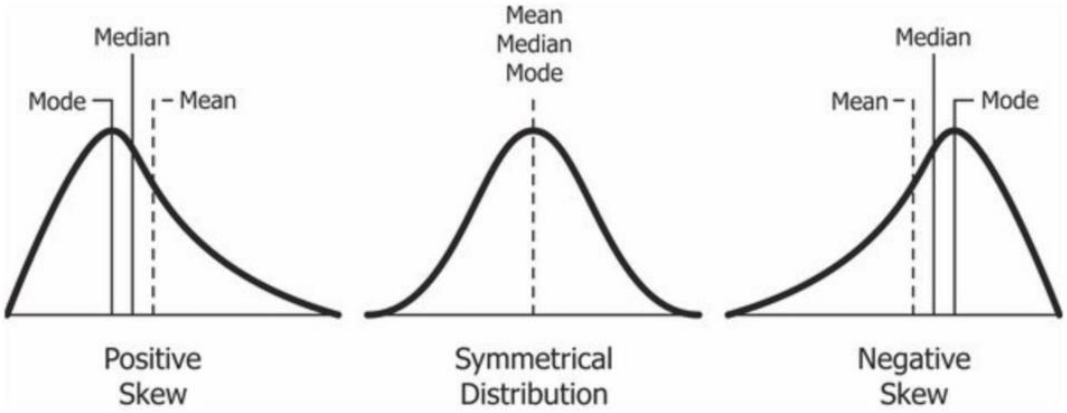


3. Fill in the missing data manually:

- Requires domain knowledge (in the field) regarding data.
- If the data shows a clear trend, this method is appropriate to use.
- Estimate the missing value based on the before & after values.

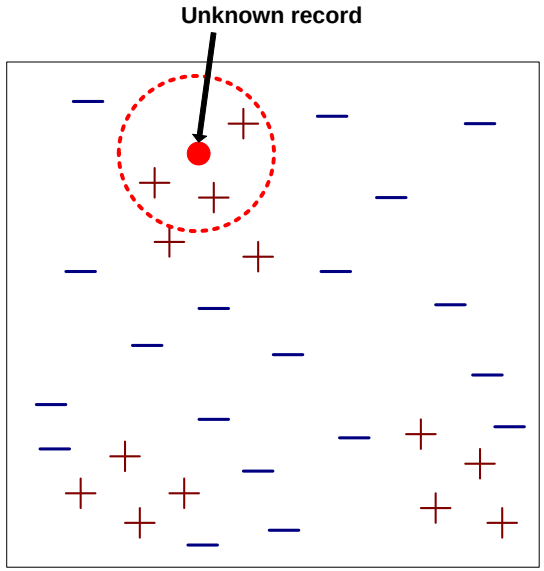
4. Use a centralized metric measure as an estimate of the missing data within the same attribute:

- For non-numeric variable data: mode value can be used.
- For symmetric distribution of numerical data: the mean value can be used.
- For biased or asymmetric distribution data: the median value can be used.



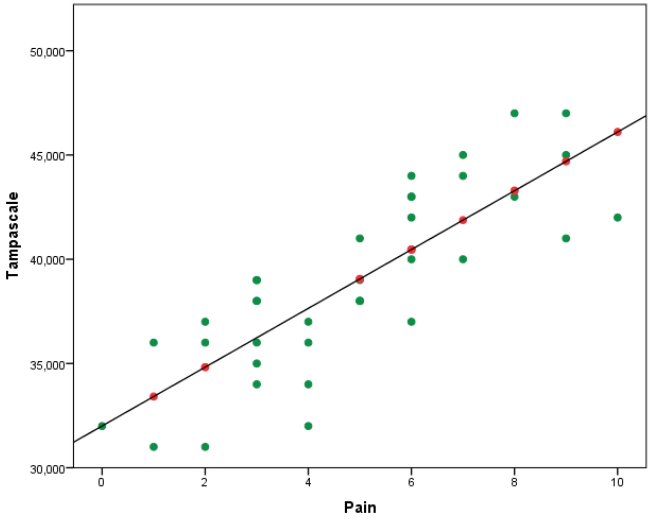
5. Use k-nearest neighbor information as an estimate of missing data:

- Identify the k-points of observation that are closest to the missing data.
- Use information for nearest neighbors as an estimate of missing data.



6. Missing data estimate through various methods of statistical imputation:

- Single imputation, multiple imputation.
- Based on regression methods, Mean Predictor Matching, Bayesian, Multivariate, etc.
- These methods can be execute using some R packages.



MISSING DATA ESTIMATE USING R:

- There are several packages in R that can be used to deal with missing data. Among them are:

i) mice

ii) Amelia

iii) missForest

iv) Hmisc

v) mi

DEALING WITH MISSING DATA THROUGH MICE PACKAGE:

- mice refer to "*Multivariate Imputation via Chained Equations*".
- It performs multiple imputations technique to estimate missing data.

- Suppose we have a variables $X_1, X_2....X_k$.
- If X_1 has missing data, then the variables $X_2, X_3....X_k$ will be used in the predictor model to estimate the missing values in X_1.
- Next, the missing data in X_1 will be replaced with the estimated values which obtained from the model.
- Similarly, if X_2, has missing data, then the variables $X_1, X_3....X_k$ will be used in the predictor model to estimate the missing values in X_2.
- And so on.

- Among the predictor models used in the mice package are:

i) PMM (Predictive Mean Matching): for numerical variable.

ii) logreg(Logistic Regression): for binary variable with 2 level.

iii) polyreg(Bayesian polytomous regression): for factor variable with >= 2 level.

iv) Proportional odds model: for ordered categorical variable >= 2 level.

v) And many more.

please refer to Package 'mice' - R Project to learn about the various models in the mice package

DEALING WITH MISSING DATA THROUGH AMELIA PACKAGE:

- The package is named after pioneer and author of American Aviation, Amelia Earhart.
- Amelia was the first female pilot that attempt to create a record as the first woman to fly solo around the world through the Atlantic Ocean in 1932, but she missing during the flight.
- There is no evidence whether she is alive or dead.

- This package uses the bootstrapping method and Expectation-Maximization algorithm to estimate the missing data in the data set.
- Step 1: Bootstrapping technique.

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

Row num	Var1	Var2	Var3
1	9	3	4
2	4	?	2
3	?	2	?

- Step 2: Imputations based on Maximization-Expectation algorithm.

Please Refer to Package 'Amelia' - R to learn more about the rules in the Amelia package.###

#imputation	Row num	Var 1	Var 2	Var 3
1	1	9	3	4
1	2	4	3	2
1	3	4	2	5
2	1	9	3	4
2	2	4	4	2
2	3	2	2	3
3	1	9	3	4
3	2	4	2	2
3	3	2	2	4

DEALING WITH INCONSISTENT DATA :

- Inconsistent data needs to be corrected or eliminated so that it does not adversely affect the data mining process.

1. Identify inconsistent data :

- This procedure needs to be done when data is merged from multiple sources or files.
- Example: A person's name may be fully specified in one data source, while the other sources could only contain the beginning and last name.

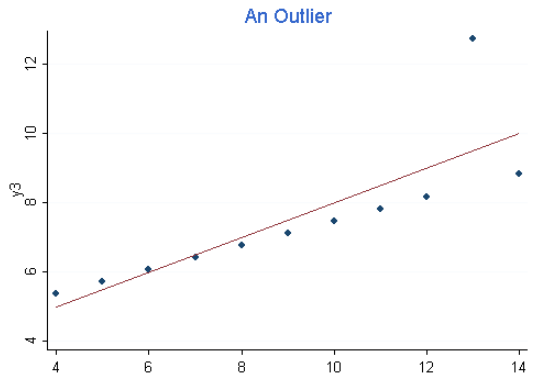
2. Domain Knowledge:

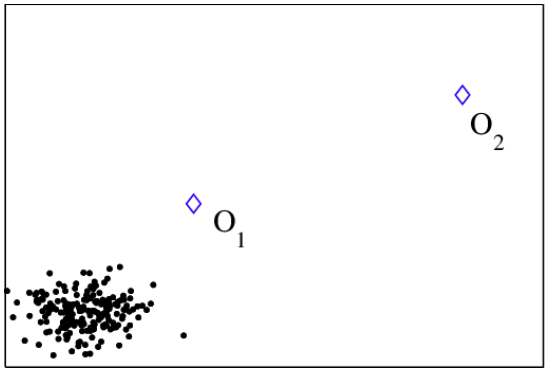
- Extensive domain knowledge in the field under study is very helpful in correcting inconsistent data.
- Example: if the district attribute is "Pasir Mas" then their State cannot be "Selangor".

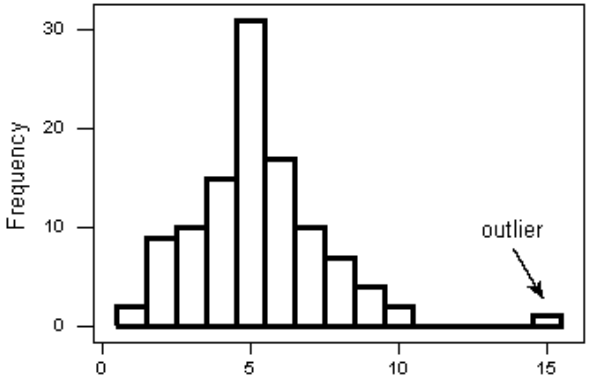
#This topic has been discussed in data integration process.

DEALING WITH OUTLIERS:

- Outliers defined as an observations located at a considerable distance from most of the data.





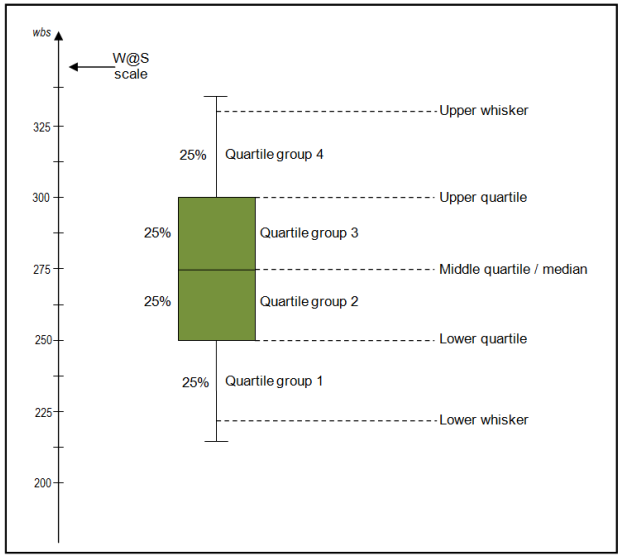


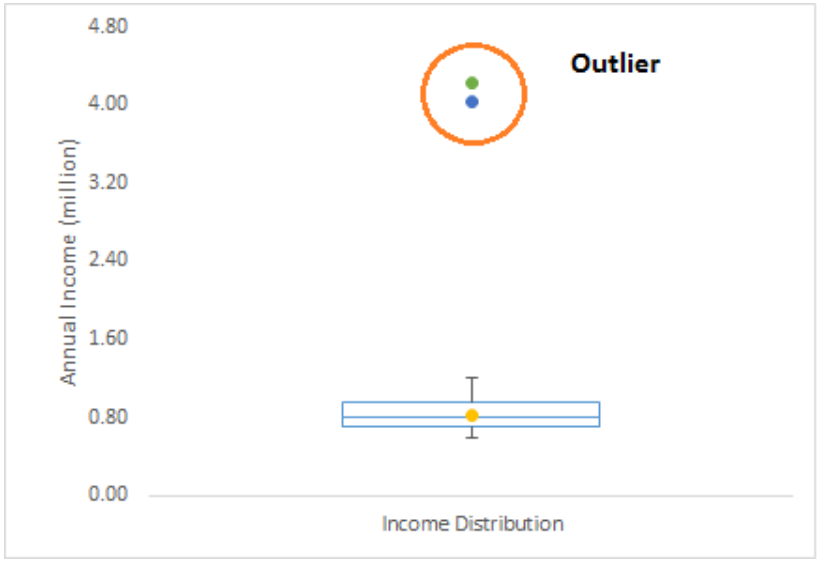
- Outliers can occur due to data entry errors, tool malfunctions or observation errors.
- However, if there are no errors in the data recording, the outliers are very important information (considered as a rare case).
- Example: Credit card fraud, massive floods, millionaire income, etc.
- Outlier can affect the accuracy of data mining if it is not identified and handled appropriately.

OUTLIERS DETECTION:

i) Univariate approach (one variable):

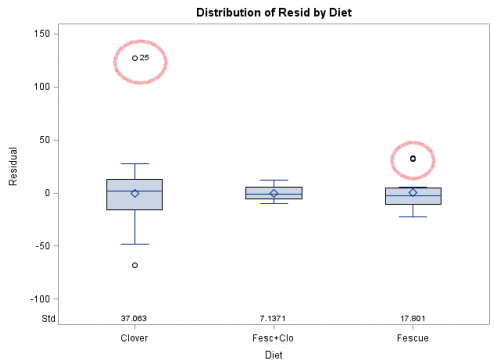
- Use boxplot method.
- For univariate continuous variables, outliers are observations that are located beyond the 1.5 x IQR (Interquartile Range).
- IQR is the difference between the 75th and 25th quartiles.

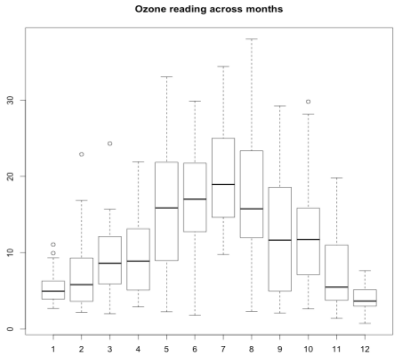


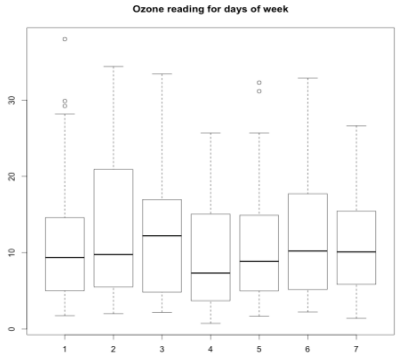


ii) Bivariate approach (two variable (X dan Y)):

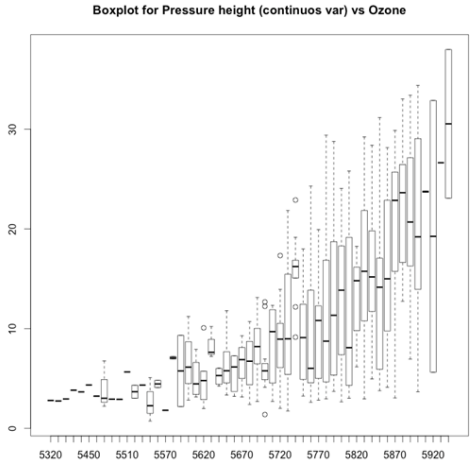
- If the variable X is categorical (level) and Y is continuous, use a boxplot.

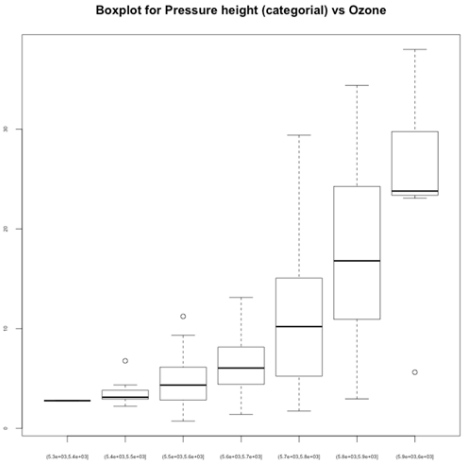




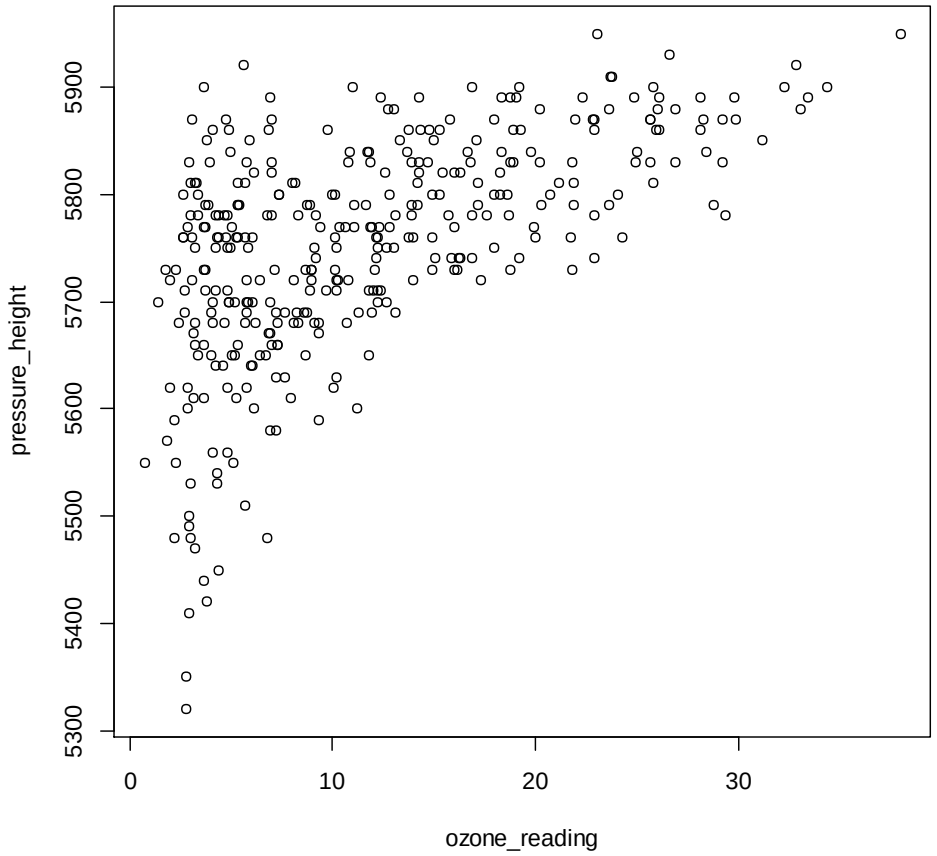


- If both X and Y are continuous, a box plot can still be used.
- However, try to transform X into a categorical form if the variation is too large





- Another approach is to use scatter plots.



- Once the outliers have been identified, the Data Scientist job is to investigate whether it is incorrect data or rare event (very important information) .

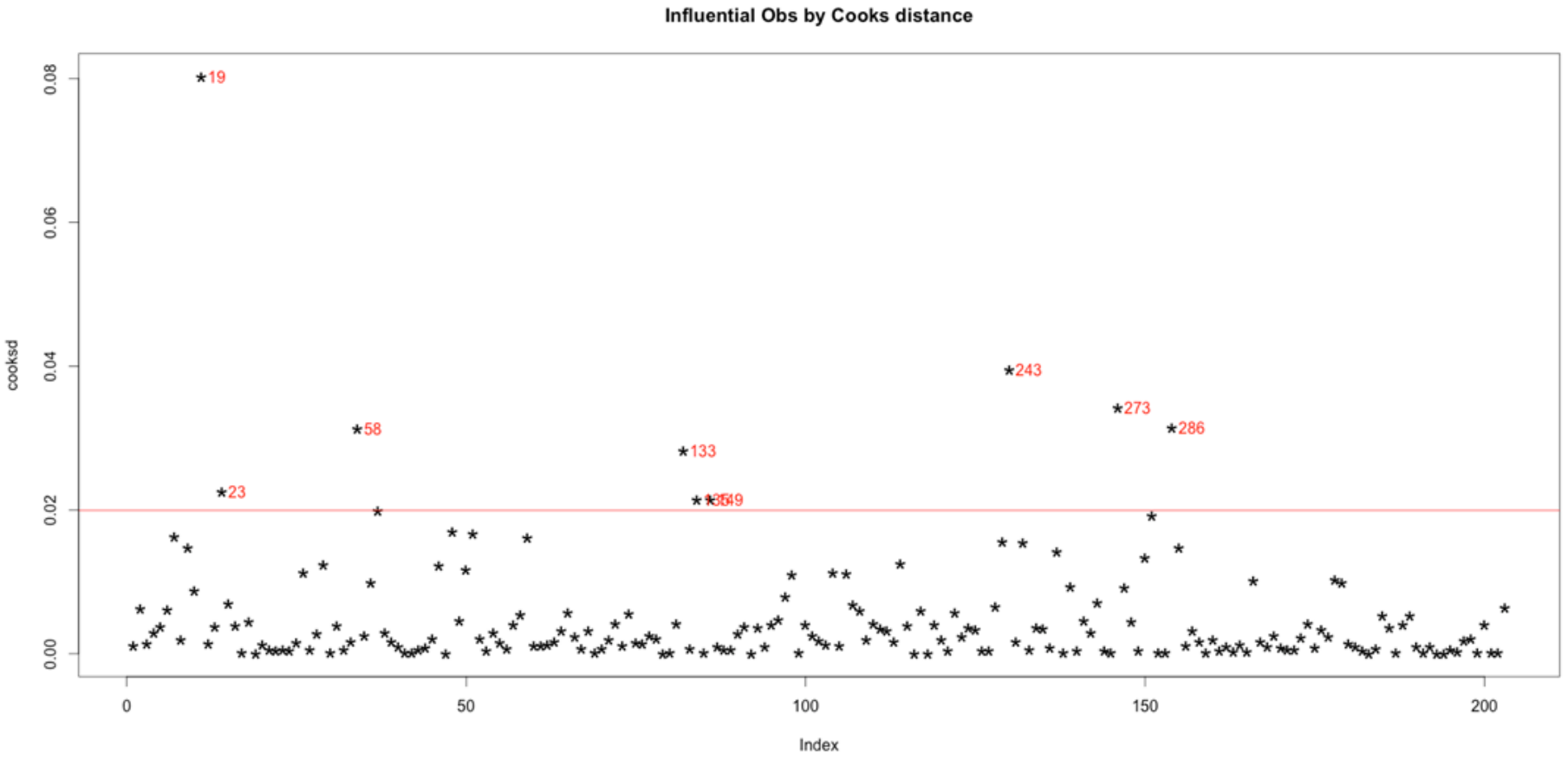
ii) Multivariat approach (supervised case):

- For data consisting of several variables, it needs to be evaluated simultaneously to determine the presence of outliers.
- To determine whether certain observations (rows) are outliers or not, the Cook-distance method can be used:

$$D_i = \frac{\sum_{j=i}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{p \times MSE}$$

- The multiple linear regression is fitted for the data with response variable Y to the explanatory variables $X_1, X_3 \ldots X_k$
- $\hat{Y}_j$ is the j-th fitted value when taking into account the values of all observations.
- $\hat{Y}_{j(i)}$ is the j-th fitted value when the observation-i is not taken into consideration.
- MSE is the mean square error.
- p is the number of regression parameter.

- In general, observations that have a Cook-distance value that is 4 times larger than the mean value of Cook-distance will be classified as influential observations.

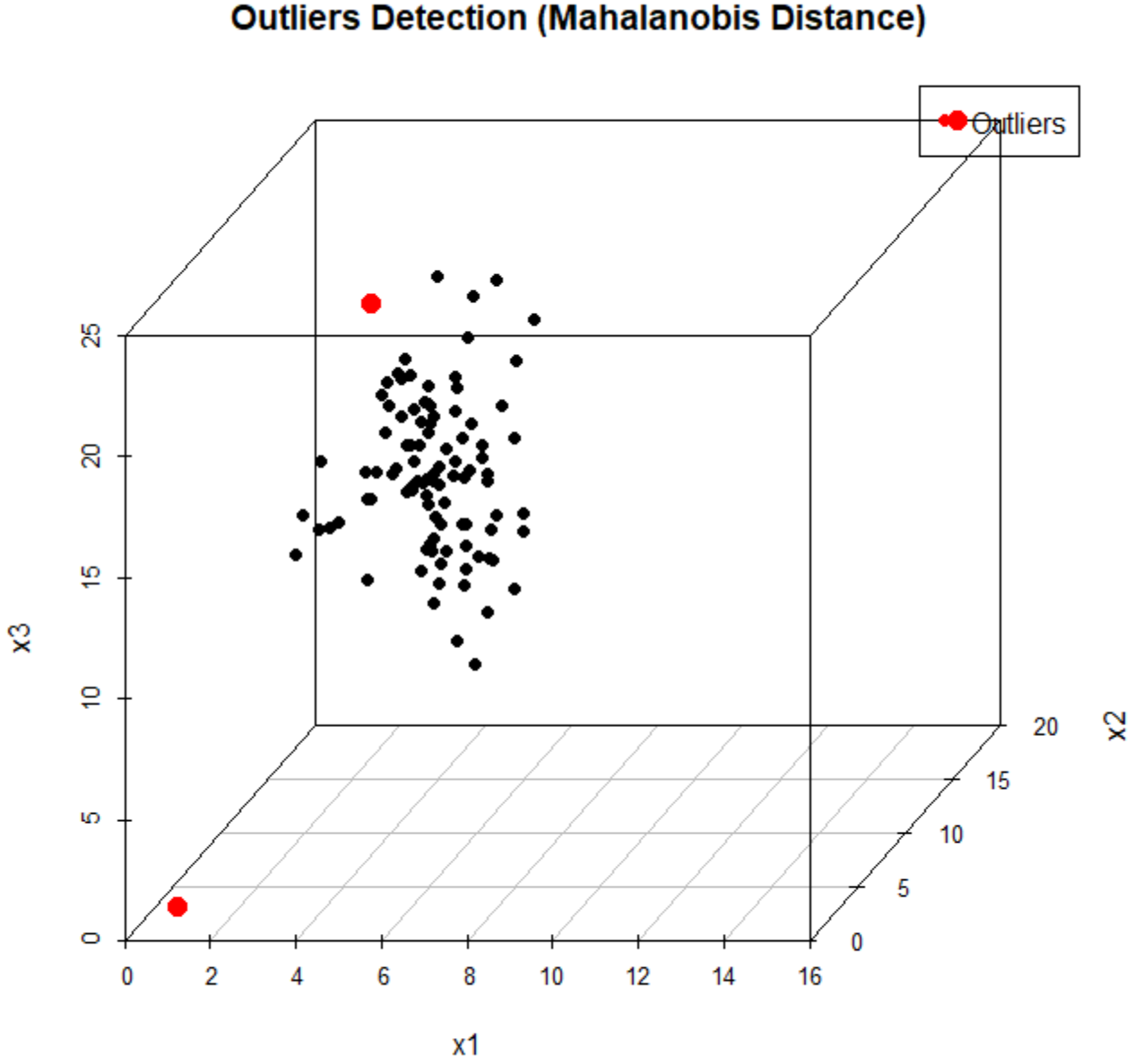


iii) Multivariat approach (unsupervised case):

- For the case of unsupervised data, distance-based methods such as the Mahalanobi distance are often used.
- The Mahalanobi distance measures which an observation differs significantly from the mean of the data in multivariate space through the following formula:

$$D_M = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- where $\mathbf{x}$ is a multivariate dataset $\left[x_1, x_2, \ldots, x_p\right]^T$.
- $\boldsymbol{\mu}$ is a mean vector.
- $\mathbf{S}$ is a covariance matrix.
- A larger Mahalanobis distance indicates that a data point is further from the mean of the data distribution.
- Outliers can be identified by setting a threshold based on the chi-square distribution.
- If D_M^2 exceeds a critical value from a chi-square distribution with k degrees of freedom (k is the number of variables), the point can be considered as an outliers.


Outliers Detection (Mahalanobis Distance)
Outliers
x3
x1
x2
0
5
10
15
20
25
0
2
4
6
8
10
12
14
16
0
5
10
15
20

TREATING OUTLIERS:

- After the outliers are identified, it needs to be treated through:

i) If the outliers are errors:

- It can be discarded or assume it as a missing data.
- Next, the value of the missing data can be predicted through the imputation method.

ii) If the outliers are really an actual data (rare events/very important information):

- It needs to be retained in the data.
- Specific statistical methods such as; Robust Statistics need to be used to analyze this type of data.
- Some data mining methods can be applied to that data that contains an outliers (Example: clustering)

REFERENCES:

- Aggarwal, C.C. (2015). *Data Mining. The Textbook*. Springer, New York.
- García, S., Luengo, J., Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer, New York.
- Jafari, R. (2022). *Hands-On Data Preprocessing in Python: Learn how to effectively prepare data for successful data analytics 1st Edition*. Packt Publishing
- Kuhn, M., Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall.
- Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., Herrera, F. (2020). *Big Data Preprocessing*. Springer, Switzerland.

NEXT TOPIC:

Data Transformation and Discretization