FIT5196 DATA WRANGLING

Week 8

Data Cleansing

By Jackie Rong

Faculty of Information Technology

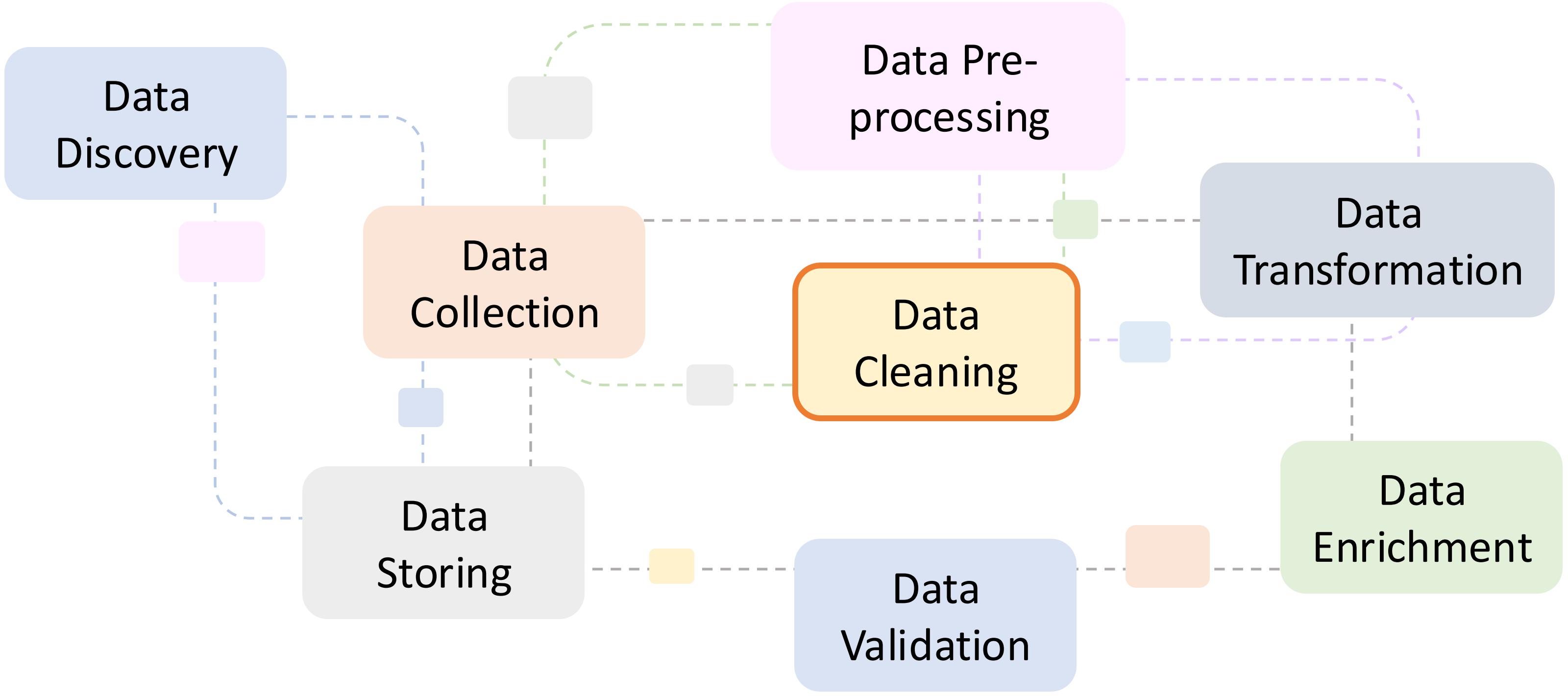
Monash University

Data Quality

- **Data quality** refers to the condition or state of data based on factors that influence its accuracy, completeness, reliability, relevance, and timeliness.
- High-quality data is essential for businesses, governments, and organizations to make informed decisions, improve operational efficiency, and gain competitive advantage.

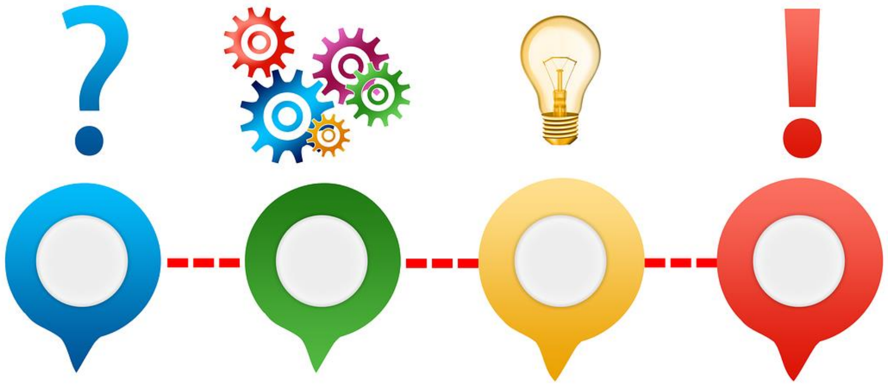
Data Wrangling Tasks (Recap)

In the **Data Pre-processing** stage, preliminary data preparation tasks are performed to make raw data more suitable for analysis.



Data Cleansing

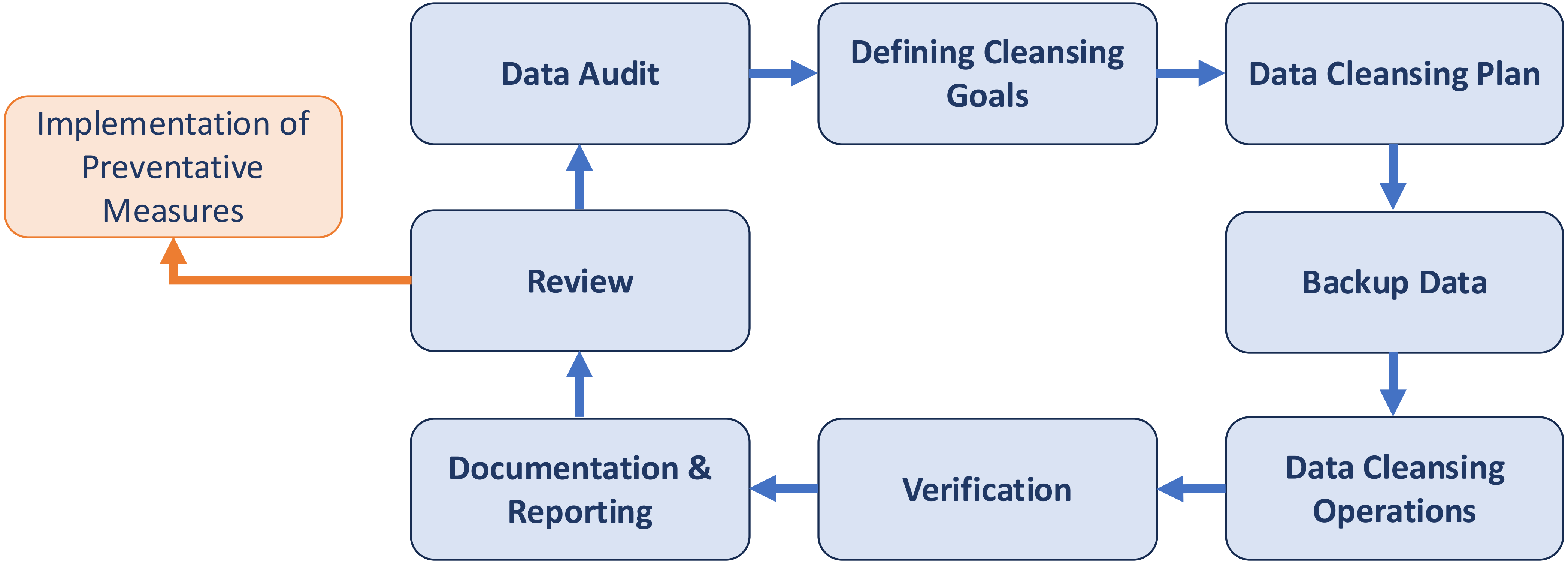
- Overview of Data Cleansing
- Data Cleansing Operations & Methods
  - Missing Data
  - Outliers



Data Cleansing

- **Data cleansing**, also known as data cleaning, is a fundamental aspect of data wrangling.
- Data cleansing involves detecting and correcting (or removing) corrupt or inaccurate records from a dataset.
- Data cleansing is a critical component of data wrangling, particularly in the era of big data, where organizations depend heavily on accurate and reliable data for making informed decisions.
- Clean data is crucial for the effectiveness of machine learning models, statistical analyses, and business intelligence tools.
- It helps in reducing errors, improving efficiency, and ultimately leading to more trustworthy insights and decisions.
- Effective data cleansing requires a mix of automated tools and human judgment, especially in complex scenarios where context and domain knowledge are vital for interpreting data accurately.

Data Cleansing Process



Data Audit

- A **data audit** is a comprehensive review of an organization's data to ensure accuracy, completeness, consistency, and reliability.
- It's a critical first step in the data cleansing process, serving as a diagnostic phase that identifies issues affecting data quality.
- A thorough data audit not only uncovers problems but also helps in understanding the overall health of the data, setting the stage for effective data management strategies.

Data Audit

Data Audit

- **Objectives of a Data Audit**
  - **Identify Data Quality Issues**
    - Discover inaccuracies, inconsistencies, duplicates, missing values, and other anomalies that compromise data integrity.
  - **Assess Data Completeness**
    - Determine if critical data is missing or incomplete.
  - **Evaluate Data Consistency**
    - Ensure data is consistent across different sources and systems.
  - **Understand Data Usage**
    - Identify how data is being used across the organization and whether it meets the needs of its users.
  - **Compliance Check**
    - Verify that the data management practices comply with relevant data protection and privacy regulations.

Data Audit

- **Data audit methods**
  - Establish metrics for measuring data quality, including accuracy, completeness, consistency, and reliability.
  - Create a map of where data resides across the organisation.
  - Engage with stakeholders to understand their data requirements, challenges they face with the current data and their expectations from the audit.
  - Select a representative sample of data for detailed analysis, rather than examining the entire dataset.
  - Use software tools to automatically scan data for common issues.
  - Review the data manually for complex or critical cases to identify issues that automated tools cannot detect.

Data Audit

- Popular data audit software and tools

talend

IBM Infosphere

ataccama

tableau Prep

Google Cloud Dataprep

Informatica DATA QUALITY

sas Data Management

Microsoft SQL Server 2012 Data Quality Services

alteryx

Data Ladder
Get the most out of your data

Defining Cleansing Goals

- **Defining cleansing goals** is a critical step in the data cleansing process, setting clear objectives for what the cleansing efforts aim to achieve.
- This step involves specifying the standards and metrics that will guide the cleaning operations and ultimately determine their success.
  - **Understanding Business Requirements**
  - **Identifying Data Quality Dimensions**
  - **Setting Specific, Measurable Goals**
  - **Prioritizing Goals**
  - **Creating a Roadmap**
  - **Continuous Improvement**
  - **Communication and Documentation**

Defining Cleansing Goals

Data Cleansing Plan

- **Creating a data cleansing plan** is a structured approach to improving the quality of data in a database, dataset, or an information system.
- A well-constructed plan ensures that data cleansing efforts are effective, efficient, and aligned with the strategic needs of the organization.

Data Cleansing Plan

1. Assess Data Quality
2. Define Cleansin Objective
3. Develop Cleansing Strategies
4. Select Tools & Resources
5. Create an Implementatic Timeline
6. Establish Monitoring & Maintenance Procedures
7. Documentation & Training

Backup Data

- Backing up data before initiating the data cleaning process is a critical step that provides a safety net against potential data loss or corruption.
- This precautionary measure offers several strategic and operational benefits:
  - Risk mitigation
  - Data integrity assurance
  - Operational continuity
  - Flexibility in data handling
  - Confidence in data quality improvements
  - Legal and compliance safeguards

Backup Data

Data Cleansing Operations

- **Data quality problems**
  - Duplicated data records
  - Inaccurate data
  - Inconsistent data
  - Incomplete data
  - Irrelevant data

- **Data cleansing operations**
  - Removing duplicates
  - Validating and correcting errors
  - Consistency checks
  - Filling missing values
  - Handling outliers

Data Cleansing Operations

Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms

Staff_ID	First_Name	Last_Name	Level	Work_Hour
S001	John	Smith	D	6
S002	Kate	Joyce	C	8
S003	Mary	Wen	D	6
S004	Jenny	Wood	D	6
S005	Jon	Dolly	E	4
S006	Amy	Yeewood	A	10
S007	Addy	Zhang	B	9
S008	Allen	Fan	B	9
S009	James	Vu	A	10
S010	Anddy	Lee	D	5
S011	Jane	Jones	C	8
S012	Mike	Giacometti	C	8
S013	Anna	Nord	E	4
S014	Sunny	Johnson	E	4
S015	Ross	Hart	A	10
S006	Amy	Yeewood	A	10
S003	Mary	Wen	D	6

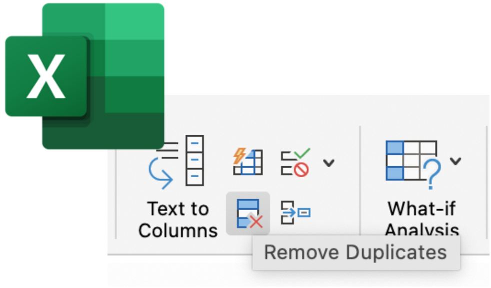
Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms

Staff_ID	First_Name	Last_Name	Level	Work_Hour
S001	John	Smith	D	6
S002	Kate	Joyce	C	8
S003	Mary	Wen	D	6
S003	Mary	Wen	D	6
S004	Jenny	Wood	D	6
S005	Jon	Dolly	E	4
S006	Amy	Yeewood	A	10
S006	Amy	Yeewood	A	10
S007	Addy	Zhang	B	9
S008	Allen	Fan	B	9
S009	James	Vu	A	10
S010	Anddy	Lee	D	5
S011	Jane	Jones	C	8
S012	Mike	Giacometti	C	8
S013	Anna	Nord	E	4
S014	Sunny	Johnson	E	4
S015	Ross	Hart	A	10

Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms



Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms

CustomerID	FirstName	LastName	Email	SignupDate
1	John	Doe	johndoe@example.com	2021-01-01
2	Jane	Doe	janedoe@example.com	2021-02-01
3	John	Doe	johndoe@example.com	2021-01-01
4	Mike	Smith	mikesmith@example.com	2021-03-01
5	John	Doe	johndoe@example.com	2021-01-01

Identifying duplicates

```sql
SELECT FirstName, LastName, Email, SignupDate, COUNT(*)
FROM Customers
GROUP BY FirstName, LastName, Email, SignupDate
HAVING COUNT(*) > 1;
```

Reviewing duplicates

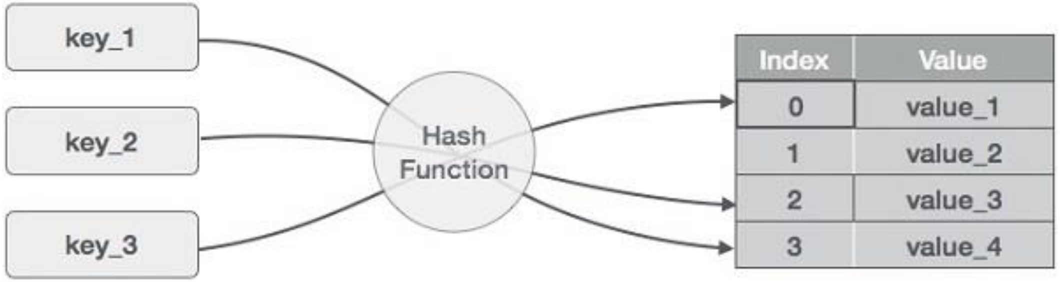
```sql
SELECT *
FROM Customers
WHERE (FirstName, LastName, Email, SignupDate) IN (
    SELECT FirstName, LastName, Email, SignupDate
    FROM Customers
    GROUP BY FirstName, LastName, Email, SignupDate
    HAVING COUNT(*) > 1
);
```

Deleting duplicates

```sql
DELETE FROM Customers
WHERE CustomerID NOT IN (
    SELECT MIN(CustomerID)
    FROM Customers
    GROUP BY FirstName, LastName, Email, SignupDate
);
```

Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms



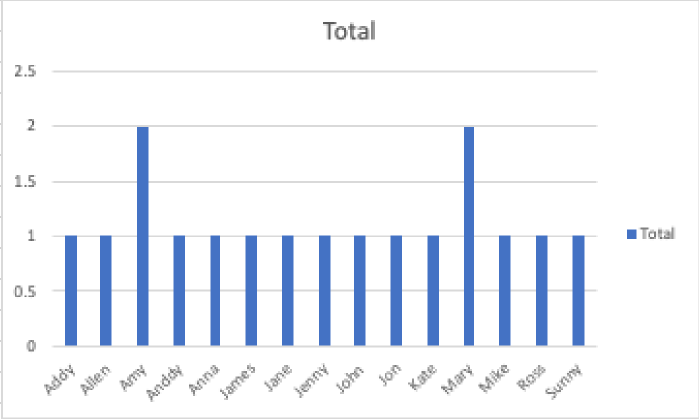
Source: https://www.tutorialspoint.com/data_structures_algorithms/hash_data_structure.htm

Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms

Staff_ID	First_Name	Last_Name	Level	Work_Hour
S001	John	Smith	D	6
S002	Kate	Joyce	C	8
S003	Mary	Wen	D	6
S003	Mary	Wen	D	6
S004	Jenny	Wood	D	6
S005	Jon	Dolly	E	4
S006	Amy	Yeewood	A	10
S006	Amy	Yeewood	A	10
S007	Addy	Zhang	B	9
S008	Allen	Fan	B	9
S009	James	Vu	A	10
S010	Anddy	Lee	D	5
S011	Jane	Jones	C	8
S012	Mike	Giacometti	C	8
S013	Anna	Nord	E	4
S014	Sunny	Johnson	E	4
S015	Ross	Hart	A	10

Row Labels	Count of Staff_ID
Addy	1
Allen	1
Amy	2
Anddy	1
Anna	1
James	1
Jane	1
Jenny	1
John	1
Jon	1
Kate	1
Mary	2
Mike	1
Ross	1
Sunny	1
Grand Total	**17**



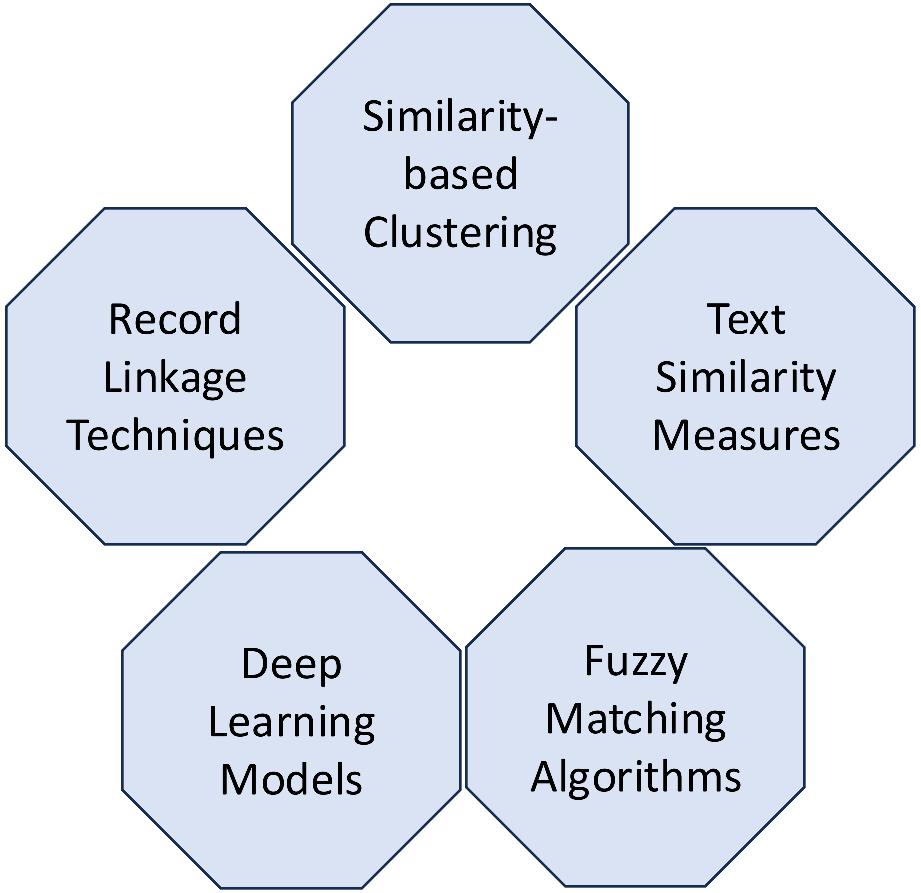
Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms

pandas

Data Cleansing Operations

- **Removing Duplicates**
  - Manual review and removal
  - Sorting and sequential check
  - Deduplication software
  - Database queries (SQL)
  - Hashing techniques
  - Pivot tables
  - Scripting and programming
  - Machine learning algorithms



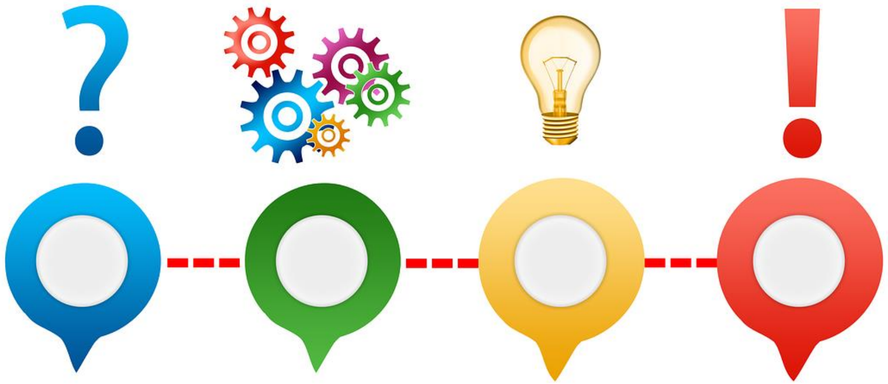
Verification

- **Verification** ensures the integrity and accuracy of data after it has been cleaned and before it is used for further processing, analysis, or decision-making.
  - **Accuracy Check**: Confirm that all data modifications (corrections, deletions, and additions) were correctly implemented.
  - **Consistency Validation**: Ensure data is consistent both internally (within the same dataset) and externally (across different data systems).
  - **Completeness Verification**: Verify that no necessary data has been inadvertently removed and that missing data issues have been suitably addressed.
  - **Quality Assurance**: Assess whether the data now meets the specified quality criteria necessary for its intended use.

Verification

Data Cleansing

- Overview of Data Cleansing
- Data Cleansing Operations & Methods
  - Missing Data
  - Outliers



Missing Data

- Reasons for missing values
  - Equipment errors
  - Absence of survey participants
  - Unavailability of GPS signals in rural areas
  - Change of circumstances, such as death, graduation, etc.
  - Filter question when a set of questions in a survey is only asked to participants who indicate they are married.

Missing Data

- Why is missing data a problem in data analysis?
  - All standard statistical methods presume complete information for all the variables included in analysis.
- **Consequences**: Ignoring or inappropriately handling missing data may lead to
  - **biased estimation**: over/under-estimated sample mean and variance
  - **Incorrect inferences/results**: garbage in garbage out

Missing Data Mechanisms

- Describe relationships between measured variables and the probability of missing data
- Deciding upon the method for analysing missing values require 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

Missing at Random (MAR)

- **MAR Definition**: the probability of missing data on a variable is related to some other measured variable (or variables) in the analysis model but not to the values of the variable itself.
  - B: a binary $n \times p$ matrix indicating the missingness of the data
  - $Y = (Y_{obs}; Y_{miss})$
    - Y_{obs}: observed part of Y
    - Y_{mis}: missing part of Y
  - η: some unknown parameter

$$p(B|Y_{obs}, Y_{miss}) = p(B|Y_{obs}, \eta)$$

  which says the probability of missingness depends on the observed portion of data Y_{obs} , and some unknown parameter η.

- **Practical issue**: no way to confirm that the probability of missing data on Y is solely a function of other measured variables.

Missing Completely at Random (MCAR)

- **MCAR Definition**: the probability of missing data on a variable is irrelated to other measured variable (or variables) and is irrelated to the values of the variable itself.
  - B: a binary $n \times p$ matrix indicating the missingness of the data
  - $Y = (Y_{obs} ; Y_{miss})$
    - Y_{obs}: observed part of Y
    - Y_{mis}: missing part of Y
  - η: some unknown parameter

$$p(B|Y_{obs}, Y_{miss}) = p(B|\eta)$$

  which says some parameter still governs the probability that R takes on a value of zero or one, but missingness is no longer related to the data.

- MCAR is a more restrictive condition than MAR.
- Both MAR and MCAR could be ignorable.

Missing Not at Random (MNAR)

- **MNAR definition**: the probability of missing data on a variable is related to the values of the variable itself, even after controlling for other variables
  - B: a binary $n \times p$ matrix indicating the missingness of the data
  - $Y \ = \ (Y_{obs} \,; Y_{miss})$
    - Y_{obs}: observed part of Y
    - Y_{mis}: missing part of Y
  - η: some unknown parameter

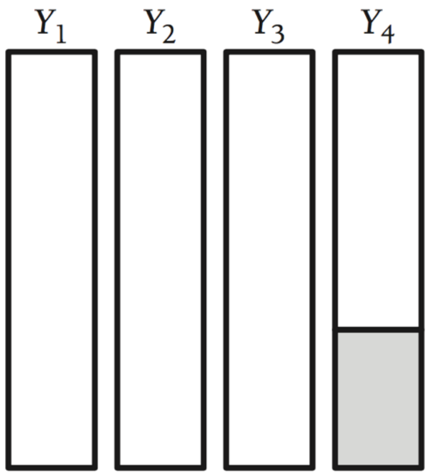
$$p(B|Y_{obs}, Y_{miss}, \eta)$$

Example

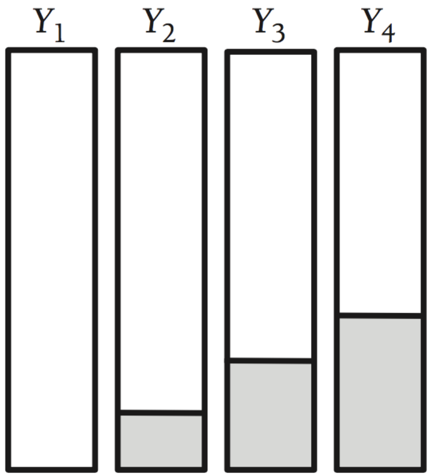
IQ	Job performance ratings			
	Complete			
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134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

Missing Data Patterns

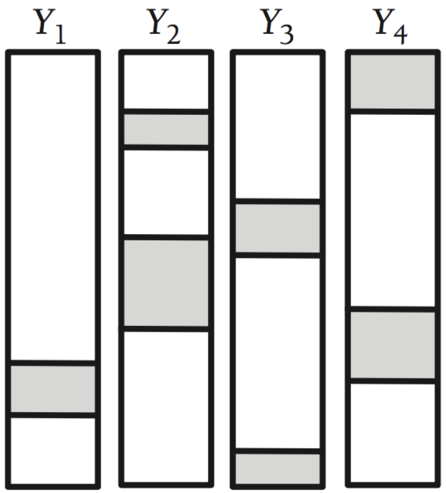
- A **missing data pattern** refers to the configuration of observed and missing values in a data set.
  - **Univariate pattern** has missing values isolated to a single variable
  - **Monotone pattern** is typically associated with a longitudinal study where participants drop out and never return.
  - **General pattern** has missing values dispersed throughout the data matrix in a haphazard fashion.



Univariate pattern

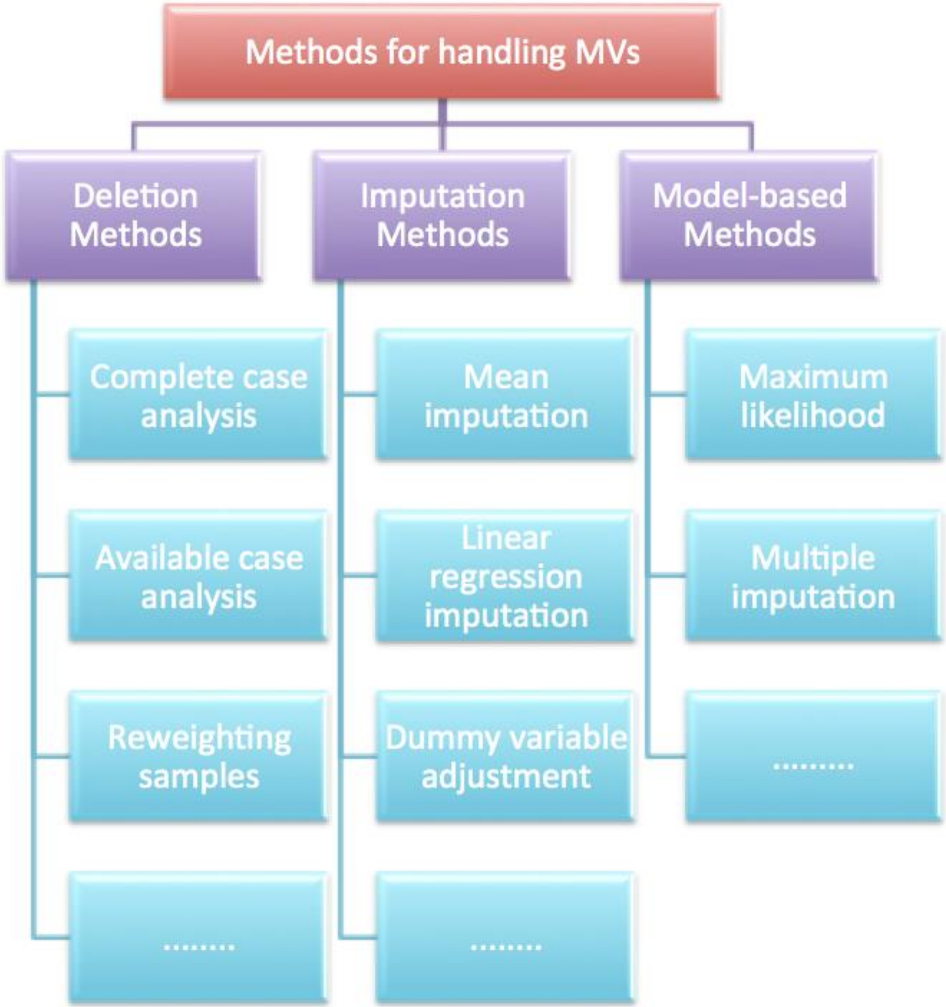


Monotone pattern



General pattern

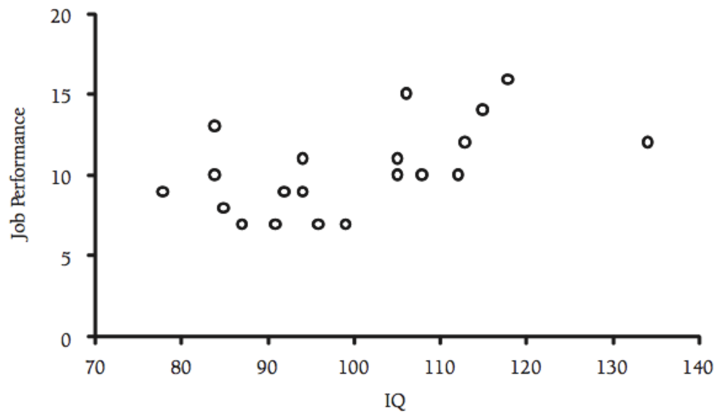
Methods for Handling Missing Data



Deletion Methods

- **List-wise Deletion** (also known as **complete-case analysis**) discards the data for any case that has one or more missing values.

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



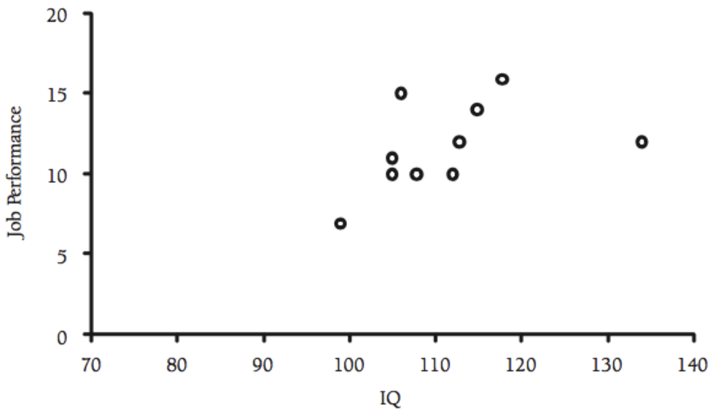


Figure is from "Applied Missing Data Analysis"

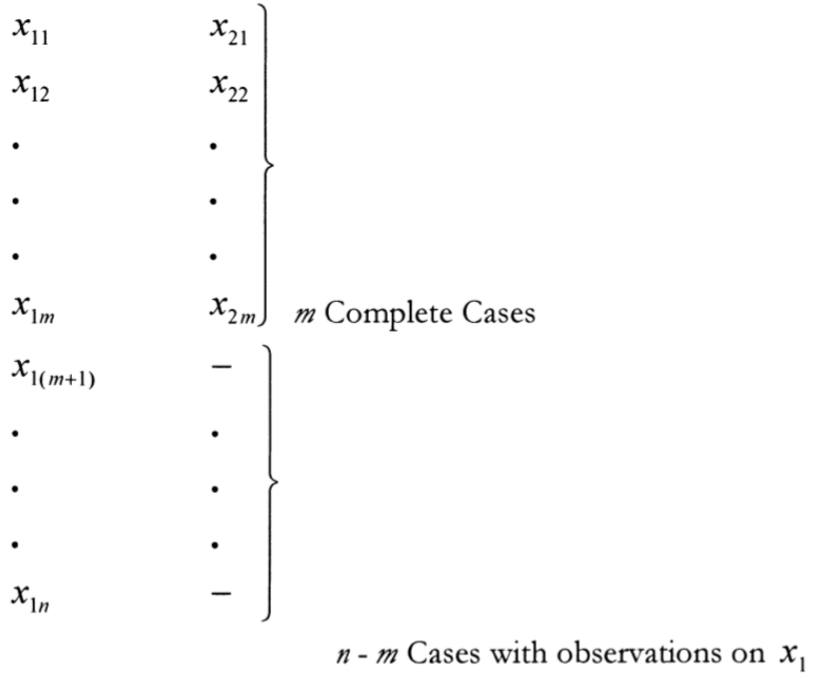
Deletion Methods

- **List-wise Deletion** (also known as **complete-case analysis**) discards the data for any case that has one or more missing values.
- Considerations
  - The primary benefit of list-wise deletion is convenience, producing a common set of cases for all analyses.
  - It assumes MCAR data and can produce distorted parameter estimates when this assumption does not hold.
  - Deleting the incomplete data records can produce a dramatic reduction in the total sample size, the magnitude of which increases as the missing data rate or number of variables increases.

Deletion Methods

- **Pairwise deletion** (also known as **available-case analysis**) attempts to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis.

Pred1	Pred2	Pred3	Pred4	outcome
5	23	34	3243	34
10		64	454	457
4.55	79			879
45.3	43	72	7623	
4.3	67	47	5489	4927
	78	56		7920
133.4	90	19	67777	
3	234	110		279
24	456	34	54389	3208



Calculate covariance

$$\bar{x}_1 = \sum_{i=1}^{n} x_{1i}$$

$$\bar{x}_2 = \sum_{i=1}^{m} x_{2i}$$

$$s_1^2 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}{n-1}$$

$$s_2^2 = \frac{\sum_{i=1}^{m} (x_{2i} - \bar{x}_2)^2}{m-1}$$

$$r_{xy}^2 = \frac{1}{m-1} \frac{\sum_{i=1}^{m} (x_{1i} - \bar{x}_{1(m)})(x_{2i} - \bar{x}_2)}{s_{1(m)}\, s_2}$$

Figure are from "A Review of Methods for Missing Data" by Therese D. Pigott

Deletion Methods

- **Pairwise deletion** (also known as **available-case analysis**) attempts to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis.
- Considerations
  - It requires MCAR data and can produce distorted parameter estimates when this assumption does not hold.
  - It is dependent on the magnitude of correlations that exist between variables.
  - It can produce estimated covariance matrices are outside of the range of -1.0 to 1.0, which causes estimation problems for multivariate analyses that use a covariance matrix as input data.
  - It is lack of a consistent sample base: cause problems in computing standard errors and covariance.

Imputation Methods

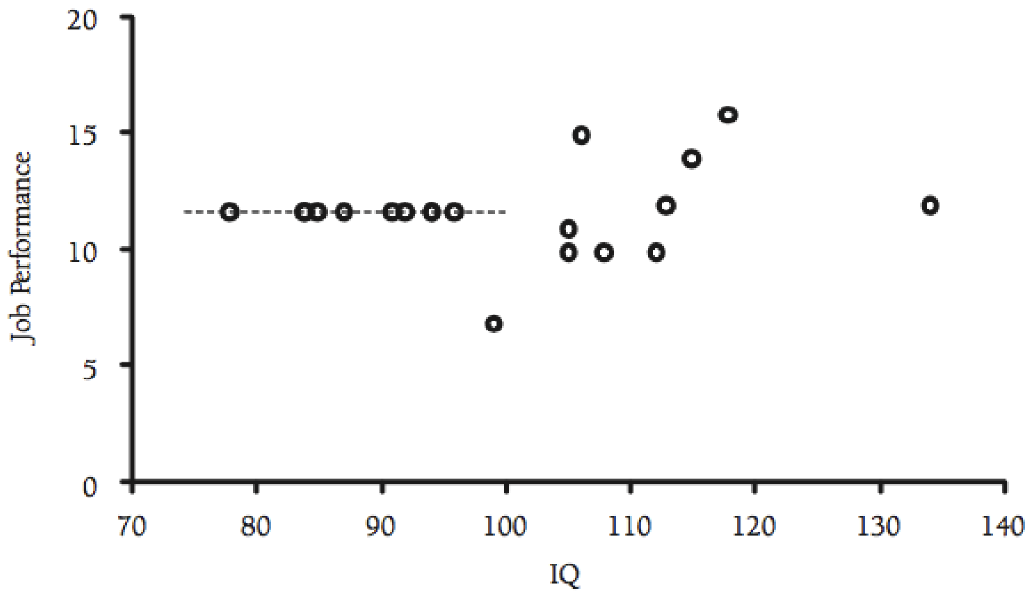
- **Single imputation**: generates a single replacement value for each missing data point.
  - Yields a complete data set
  - Produces biased parameter estimates
  - Underestimates standard errors
- Methods
  - Mean Imputation
  - Regression Imputation
  - Stochastic Regression Imputation

Imputation Methods

- **Arithmetic mean imputation** (also referred to as **mean substitution**) takes the seemingly appealing tack of filling in the missing values with the arithmetic mean of the available cases

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

$$\mu_{complete} = 10.35, \ \mu_{miss} = 11.7, \ \mu_{impute} = 11.7$$



Imputation Methods

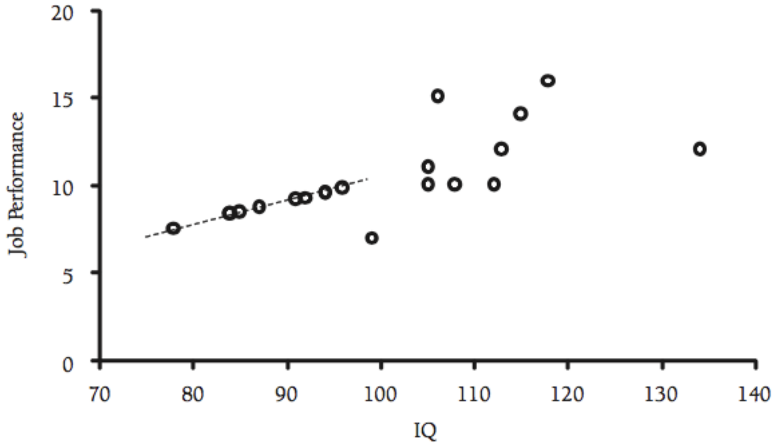
- **Regression imputation** replaces missing values with predicted scores from a regression equation.
  - Basic idea: use information from the complete variables to fill in the incomplete variables.
  - Two steps:
    1. Estimate a set of regression equations that predict the incomplete variables from the complete variables.
    2. Generate predicted values for the incomplete variables

Imputation Methods

- **Regression imputation** replaces missing values with predicted scores from a regression equation.
  - Basic idea: use information from the complete variables to fill in the incomplete variables.

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

$$
\begin{aligned}
JP_i &= \widehat{\beta_0} + \widehat{\beta_1}(IQ_i) \\
&= -2.065 + 0.123\,(IQ_i)
\end{aligned}
$$



IQ	Job performance	Predicted score
78	—	7.53
84	—	8.27
84	—	8.27
85	—	8.39
87	—	8.64
91	—	9.13
92	—	9.25
94	—	9.50
94	—	9.50
96	—	9.74
99	7	—
105	10	—
105	11	—
106	15	—
108	10	—
112	10	—
113	12	—
115	14	—
118	16	—
134	12	—

Imputation Methods

- **Stochastic regression imputation** adds random residuals to the predicate values generated by standard regression imputation.
  - Basic idea: to restore lost variability to the data and effectively eliminate the biases associated with standard regression imputation methods.
  - Three steps:
    1. Estimate a set of regression equations that predict the incomplete variables from the complete variables.
    2. Generate predicted values for the incomplete variables
    3. Add a normally distributed residual term to each predicted score

Imputation Methods

- **Stochastic regression imputation** add random residuals to the predicate values generated by standard regression imputation.
  - Basic idea: to restore lost variability to the data and effectively eliminate the biases associated with standard regression imputation methods.

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

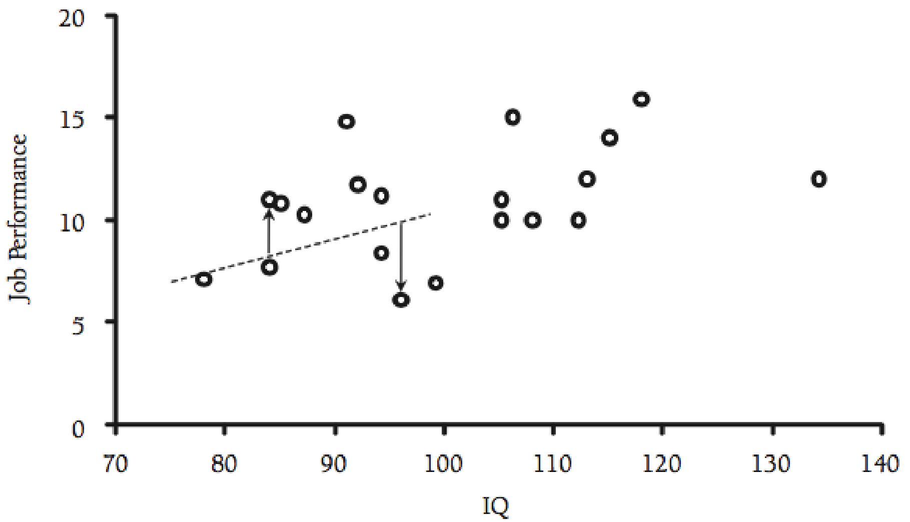
$$P_i = \widehat{\beta_0} + \widehat{\beta_1}(IQ_i) + z_i$$
$$= -2.065 + 0.123\,(IQ_i) + z_i$$

and $z_i \sim Normal(0, \sigma^2{}_{JP|IQ})$
where $\sigma^2{}_{JP|IQ}$ is the residual variance

Imputation Methods

- **Stochastic regression imputation** add random residuals to the predicate values generated by standard regression imputation.
  - Basic idea: to restore lost variability to the data and effectively eliminate the biases associated with standard regression imputation methods.

IQ	Job performance	Predicted score	Random residual	Stochastic imputation
78	—	7.53	–0.35	7.18
84	—	8.27	2.70	10.97
84	—	8.27	–0.59	7.68
85	—	8.39	2.39	10.78
87	—	8.64	1.64	10.28
91	—	9.13	5.77	14.90
92	—	9.25	2.47	11.72
94	—	9.50	–1.04	8.46
94	—	9.50	1.69	11.19
96	—	9.74	–3.58	6.16
99	7	—	—	—
105	10	—	—	—
105	11	—	—	—
106	15	—	—	—
108	10	—	—	—
112	10	—	—	—
113	12	—	—	—
115	14	—	—	—
118	16	—	—	—
134	12	—	—	—



The only procedure in this chapter that gives unbiased parameter estimates under an MAR missing data mechanism.

Evaluate a missing-data method

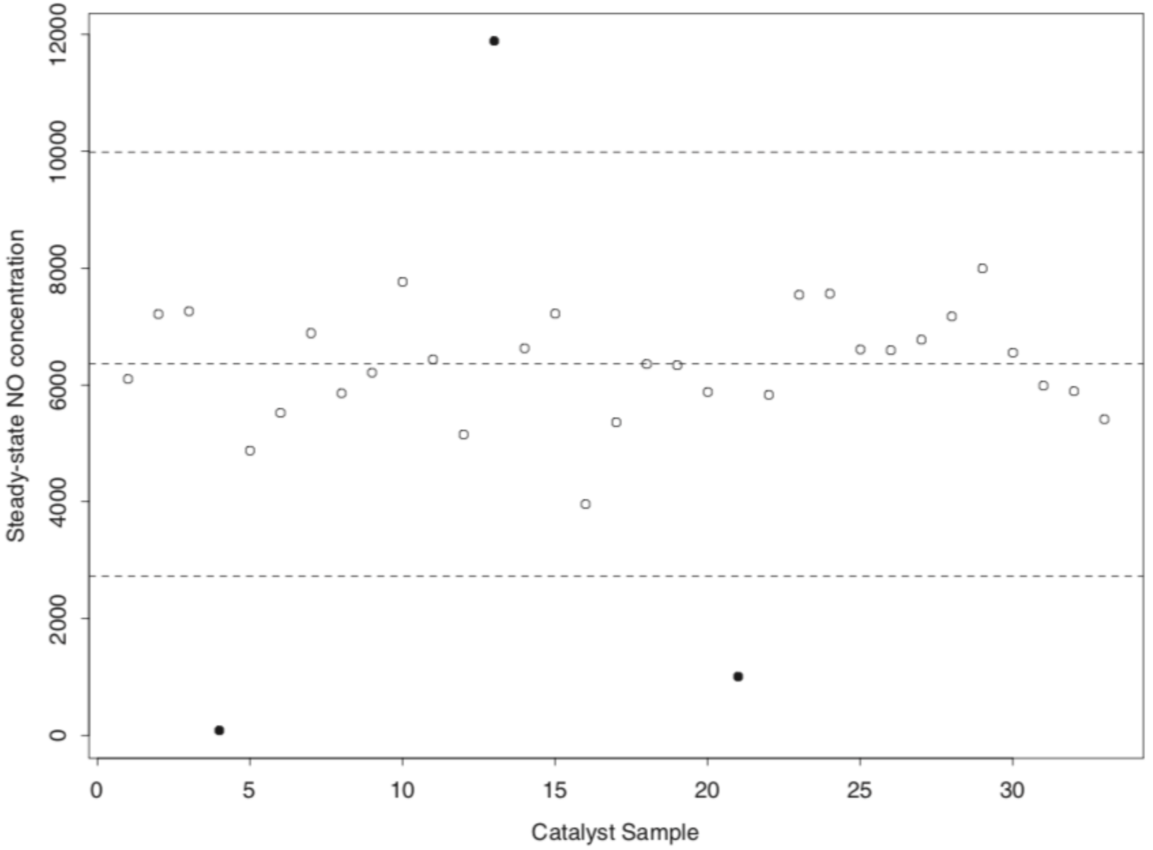
- **Minimise bias**
  - Although it is well-known that missing data can introduce bias into parameter estimates, a good method should make that bias as small as possible.
- **Maximise the use of available information**
  - We want to avoid discarding any data, and we want to use the available data to produce parameter estimates that are efficient (i.e., have minimum sampling variability).
- **Yield good estimates of uncertainty**
  - We want accurate estimates of standard errors, confidence intervals and p-values.

Imputation Methods

- **Single imputation**: generates a single replacement value for each missing data point.
  - Yields a complete data set
  - Produces biased parameter estimates
  - Underestimates standard errors
- Methods
  - Mean Imputation
  - Regression Imputation
  - Stochastic Regression Imputation

Outliers

- 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.[1]
- An **outlier** is a data point that appears to be inconsistent with the nominal behaviour exhibited by most of the other data points in a specified collection.



[1]Hawkins, D. 1980. Identification of Outliers. Chapman and Hall.

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.

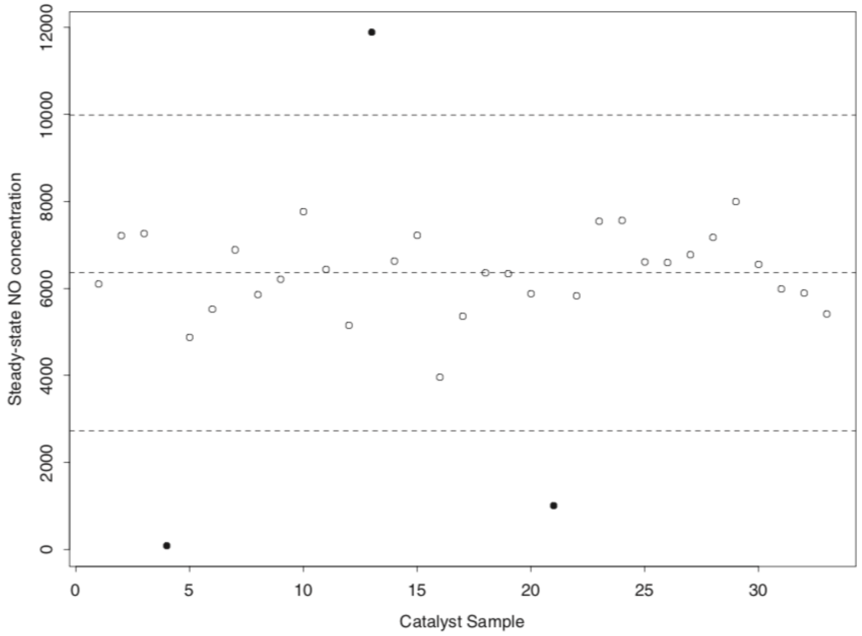
Impacts of Outliers

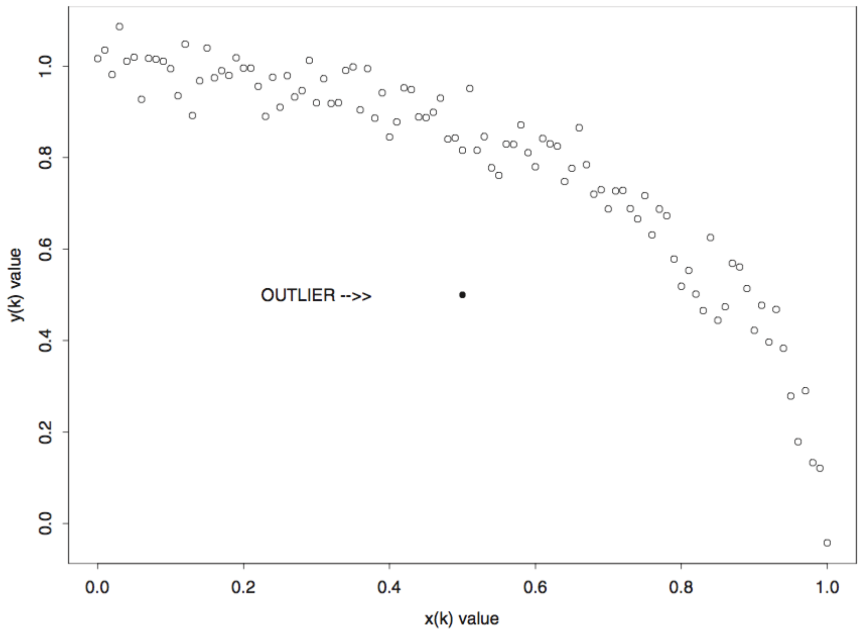
- Outliers can increase the error variance and reduces the power of statistical tests.
- If the outliers are non-randomly distributed, they can decrease normality.
- Outliers can bias or influence estimates that may be of substantive interest.
- Outliers can also impact the basic assumption of Regression,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

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.





Univariate Outlier Detection

- 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.

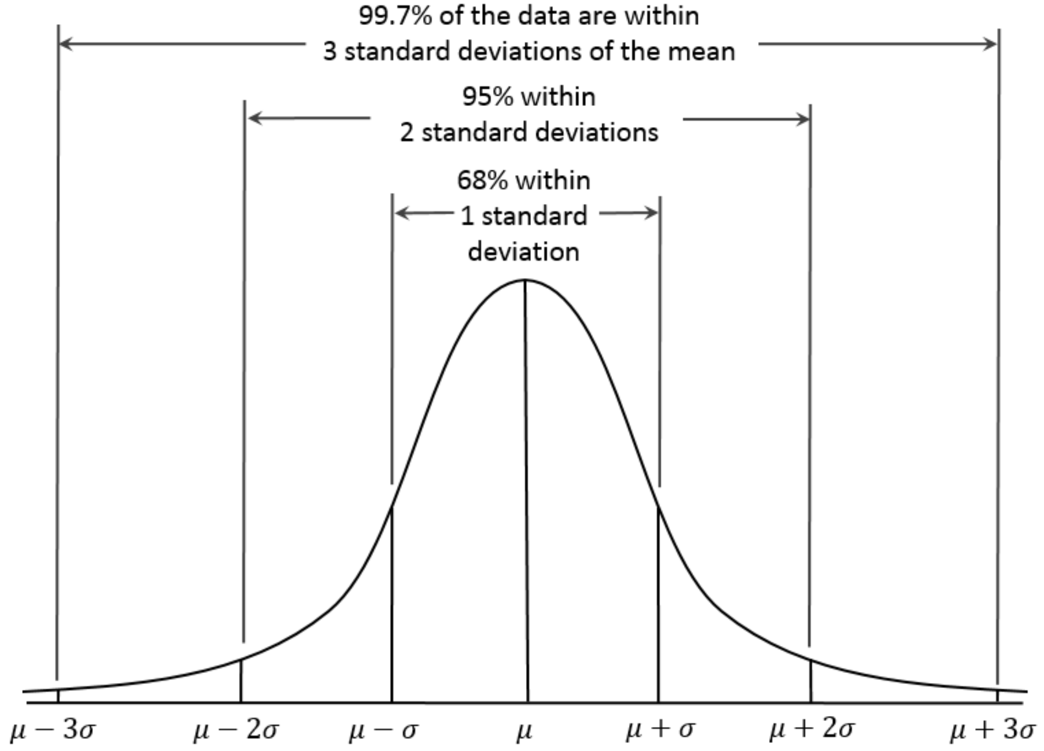
- Questions
  - How do we define the nominal data reference value x_0?
  - How do we define the scale of natural variation ζ?
  - How do we choose the threshold parameter t?

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 Independent and identically distributed 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

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

As 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
  - $x_0 \ = x^\dagger$
  - $\zeta = \ S$ = $1{:}\,4826 \times \ median\{|x_k - x^\dagger|\}$
  - x_k is an outlier if

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

- 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

- 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

- 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

$$x_0 = x^{\dagger}$$
$$\zeta = IQR$$

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

Quartile-based Detection and Boxplots

- Symmetric boxplot rule

$$|x_k - x^{\dagger}| > t \times IQR$$

- Asymmetric boxplot rule

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

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

Multivariate Outlier Detection

- 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

$$\begin{aligned}\sum_j \epsilon_j^2 &= \sum_j ((\sum_{i=1}^{d} w_i x_{j,i} + w_{d+1}) - y_j)^2 \\ &= ||\mathbf{Dw} - \mathbf{y}||^2\end{aligned}$$

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

- Closed form solution

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

Linear Models

- Regression with and without outliers

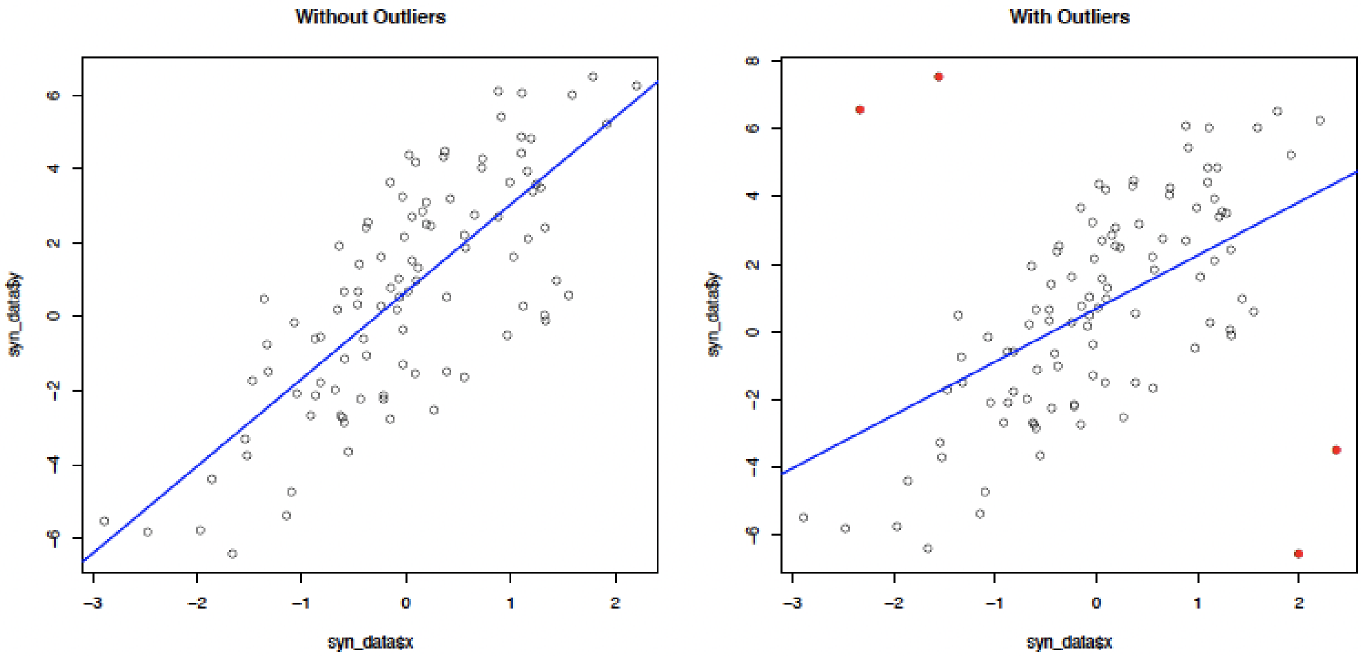
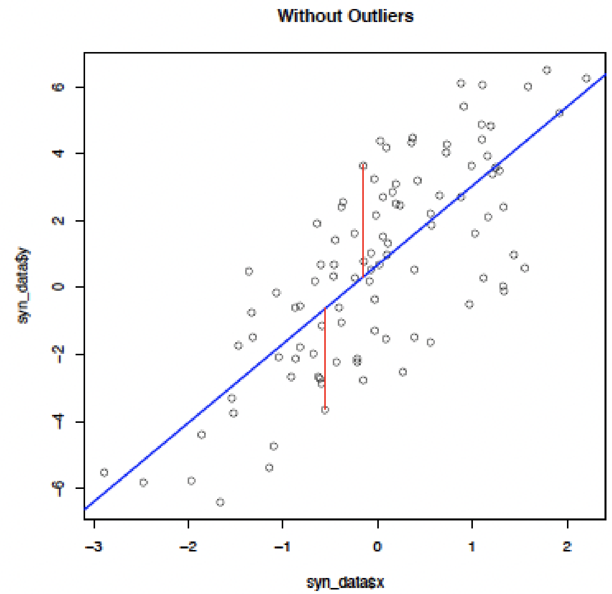


Figure : $\mathbf{y} = 2\mathbf{x} + 0.5 + \epsilon$

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.



Summary & To-do List

- Review content in Week 8.

- Assessments
  - Form your group for Assessment 2 from the same applied session in Allocate+
  - Assessment 2 specification will be released soon
  - Read the tasks in Assessment 2 and start to allocate tasks.

- Next week: Data Transformation