Data preparation and preprocessing :
Missing value management

Data Cleaning

- Data cleaning tasks:

  - **Fill in missing values**

  - Identify outliers and smooth out noisy data

  - Correct inconsistent data

  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

- Missing data may be due to

  - equipment malfunction

  - inconsistent with other recorded data and thus deleted

  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry

  - not register history or changes of the data

- Missing data may need to be inferred.

Effect of Noisy data on result accuracy

age	income	student	buys_computer
<=30	high	no	?
>40	medium	yes	?
31...40	medium	yes	?

Effect of Noisy data on result accuracy
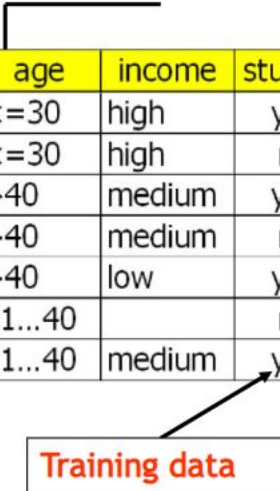
age	income	student	buys_computer
<=30	high	no	?
>40	medium	yes	?
31...40	medium	yes	?

Testing data or actual data

Effect of Noisy data on result accuracy

age	income	student	buys_computer
<=30	high	yes	**yes**
<=30	high	no	**yes**
>40	medium	yes	**no**
>40	medium	no	**no**
>40	low	yes	**yes**
31...40		no	**yes**
31...40	medium	yes	**yes**

age	income	student	buys_computer
<=30	high	no	?
>40	medium	yes	?
31...40	medium	yes	?

Testing data or actual data

Effect of Noisy data on result accuracy

age	income	student	buys_computer
<=30	high	yes	**yes**
<=30	high	no	**yes**
>40	medium	yes	**no**
>40	medium	no	**no**
>40	low	yes	**yes**
31...40		no	**yes**
31...40	medium	yes	**yes**

Training data

age	income	student	buys_computer
<=30	high	no	**?**
>40	medium	yes	**?**
31...40	medium	yes	**?**

Testing data or actual data

Effect of Noisy data on result accuracy

age	income	student	buys_computer
<=30	high	yes	yes
<=30	high	no	yes
>40	medium	yes	no
>40	medium	no	no
>40	low	yes	yes
31...40		no	yes
31...40	medium	yes	yes

Training data

Data Mining

Discover only those rules which contain support (frequency) greater >= 2

age	income	student	buys_computer
<=30	high	no	?
>40	medium	yes	?
31...40	medium	yes	?

Testing data or actual data

Effect of Noisy data on result accuracy

age	income	student	buys_computer
<=30	high	yes	**yes**
<=30	high	no	**yes**
>40	medium	yes	**no**
>40	medium	no	**no**
>40	low	yes	yes
31...40		no	yes
31...40	medium	yes	yes

Training data

Data Mining

Discover only those rules which contain support (frequency) greater >= 2

- If 'age <= 30' and income = 'high'   then buys_computer = 'yes'

- If 'age > 40' and income = 'medium' then buys_computer = 'no'

Testing data or actual data

age	income	student	buys_computer
<=30	high	no	?
>40	medium	yes	?
31...40	medium	yes	?

Effect of Missing data on result accuracy

age	income	student	buys_computer
<=30	high	yes	**yes**
<=30	high	no	**yes**
>40	medium	yes	**no**
>40	medium	no	**no**
>40	low	yes	yes
31...40		no	yes
31...40	medium	yes	yes

Training data

Data Mining

Discover only those rules which contain support (frequency) greater >= 2

- If 'age <= 30' and income = 'high'   then buys_computer = 'yes'

- If 'age > 40' and income = 'medium' then buys_computer = 'no'

Due to the missing value in training dataset, the accuracy of prediction decreases and becomes "66.7%"

Testing data or actual data

age	income	student	buys_computer
<=30	high	no	?
>40	medium	yes	?
31...40	medium	yes	?

Types of Missingness

- Missing at Random (MAR)
- Missing Completely at Random (MCAR)
- Missing Not at Random (MNAR)

Missing at Random (MAR)

- MAR data
  - there is a systematic relationship between the propensity of missing values and the observed data (Other attributes)
  - but not the missing data (Same attribute)
  - missingness of data can be predicted by other features in the dataset.

# Missing at Random (MAR)

- **Mileage table**

Model	Year	Color	Mileage	Price
Chevrolet	2014	NaN	10000	50000
Ford	2001	White	NaN	20000
Toyota	2005	Red	NaN	. 30000
Crysler	2019	Black	0	100000

- the mileage has few missing values.
- the manufacturing year of cars with missing values is lower than in other examples
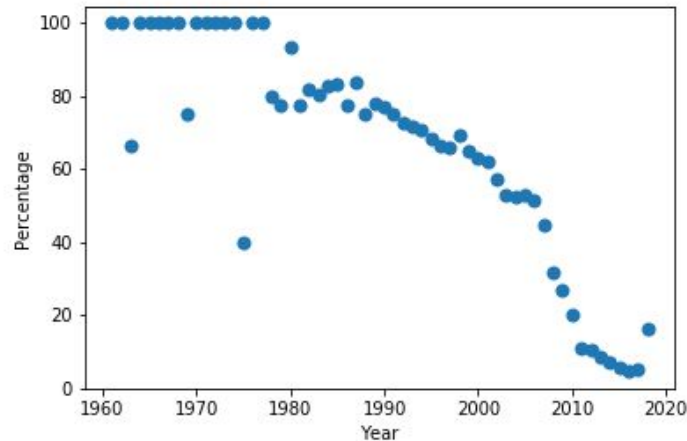
Graph for Percentage of missing values in the mileage column depending on the year of the car

Missing at Random (MAR)
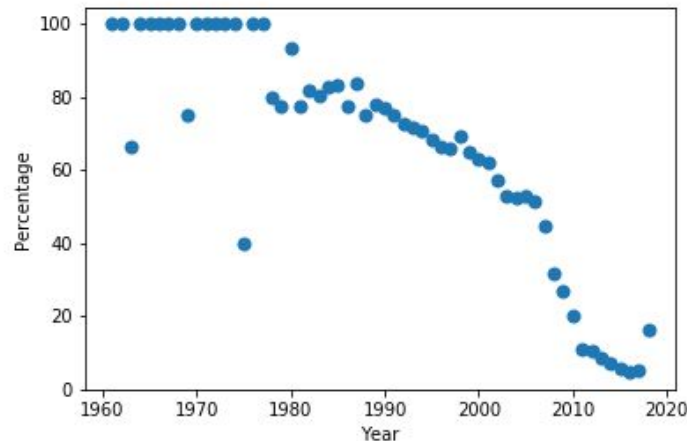
Graph for Percentage of missing values in the mileage column depending on the year of the car

- Observation
  - Clear correlation between the percentage of missing values in the mileage column and the manufacturing year of the car
  - It is clearly seen that older the car, more the probability that the mileage will not be provided by the seller of the car.
- we can predict the missingness of the mileage of the car, from its manufacturing year (may be with help of some expert)

Missing Completely at Random (MCAR)

- No relationship between the missingness of the data and any other values(observed or missing)

- Easiest to understand.

- Missing data has nothing to do with any observed/non-observed data
  - it's just missing.
  - no logic in missingness

Missing Completely at Random (MCAR)

- **Mileage table**

Model	Year	Color	Mileage	Price
Chevrolet	2014	NaN	10000	50000
Ford	2001	White	NaN	20000
Toyota	2005	Red	NaN	. 30000
Crysler	2019	Black	0	100000

- In the above given table there is missing value in the color column
- random and non systematic one.
- May be Someone just forgot to mention the color

Missing Not at Random (MNAR)

- MNAR data is the most complicated
  - one both in terms of finding it and dealing with it.
- The fact that the data is missing is related to the unobserved data,
  - i.e. the data that we don't have, the missingness is related to factors that we didn't account for.

Missing data: Another example (MAR)

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

- The missing data here is influenced only by the complete (observed) variables and not by the characteristics of the missing data itself
- IQ score is missing for youngsters (age < 44 yo)

Missing Data Another Example (MCAR)

Complete data	
Age	**IQ score**
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	**IQ score**
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

- For the given data it is MCAR
- If the data can be predicted for missing data
  - Either based on the complete variable Age (Common sense)
  - or the Missing variable IQ score (regression or any ither method)
  - **then the data is not MCAR**

Missing Data Another Example (MNAR)

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

It is impossible to detect MNAR cases without knowing the a actual missing values

How to Handle Missing Data?

1. **Ignore** the tuple (record)
   - usually done when class label is missing (assuming the tasks in classification)
   - It is not effective when the percentage of missing values per attribute varies considerably.

2. **Fill in** the missing value manually:
   - tedious + infeasible?

3. **Use a global constant** to fill in the missing value
   - (introduces a new class)
   - Normally used for categorical values
   - Not a good practice as it group different class of dat in to one new class and will lead to wrong information

How to Handle Missing Data?

4. Use the attribute values mean
   - to fill in the missing value
   - for all samples belonging to the same class to fill in the missing value:
     - smarter then **replacing by Mean** in case of classification

5. Use the most probable value to fill in the missing value

6. Use regression methods

Solution to Mising Values

- Deletion
- Imputation
  - Single Value Imputation
  - Hot/Cold Deck Imputation
  - KNN based Imputation
  - Regression based Imputation
  - Multiple imputation

Method 1 : Deletion

- **Listwise Deletion**
  - entire record is excluded from analysis if any single value is missing
  - same N (number of records) for all analysis
  - For MCAR

- **Pairwise Deletion**
  - number of records taken into consideration denoted "N" will vary according to the studied variable (column) , and for instance we could compute the mean for 2 features (Complete VS missing)
  - while dividing by the number of samples , we end up dividing by different N , one is the total number of rows and the other is the total number on complete values on the missing feature

Listwise deletion

--	3	2
8	--	2
1	5	8
1	3	5
2	4	3
4	5	3

Before Listwise deletion

1	5	8
1	3	5
2	4	3
4	5	3

After Deletion

- Whole observation that contains a missing value in any variable is discarded;
- Discarded portion of that observation will not be used when building "cross product" matrices such as the covariance or correlation matrix
- The operation used by regression procedures to deal with missing values.

Pairwise Deletion

--	3	2
8	--	2
1	5	8
1	3	5
2	4	3
4	5	3

Before Listwise deletion

1	5
1	3
2	4
4	5

After Deletion   c[1,2]

8	2
1	8
1	5
2	3
4	3

c[1,3]

3	2
5	8
3	5
4	3
5	3

c[2,3]

- If Corelation matrix C[I,j] is to be prepared to indicate corelationship between ith and jth column
  - The missing values *for each pair* of variables are deleted based on whether either variable contains a missing value.
  - The element C[1,2] is computed by using observations Row 3–6 because the first observation has a missing value for X1 and the second observation has a missing value for X2.
  - The element C[1,3] is computed by using observations 2–6 because the first observation is missing for X1.
  - The element C[2,3] is computed by using observations 1 and 3–6 because the second observation is missing for X2.

Listwise deletion (Note1)

- May lead to huge data loss and so the information loss
- Can be applied to any statistical test (SEM, multi-level regression, etc.)
- In the case of MCAR, both the parameters estimates and its standard errors are unbiased
  - As there is no dependency among attributes so deletion of some instances will not be affected
- In case of MAR and MNAR, listwise deletion can severely bias estimates of means, regression coefficients and correlations
- In the case of MAR if missing value on X is dependent on some independent variable will not lead to bias
- In the case of MAR if missing value on X is dependent on some dependent variable, then removal of the whole record will lead to bias
  - a model $y=\beta_0+\beta_1X_1+\beta_2X_2+\epsilon$ is given
  - if the probability of missing data on X1 is independent of Y will not lead to bias

Listwise deletion (Note2)

- The missing data mechanism the depends on the values of the independent variables are the same as stratified sampling. And stratified sampling does not bias your estimates

- In the case of logistic regression, if the probability of missing data on any variable depends on the value of the dependent variable, but independent of the value of the independent variables, then the listwise deletion will yield biased intercept estimate, but consistent estimates of the slope and their standard errors (Vach 1994)

- However, logistic regression will still fail if the probability of missing data is dependent on both the value of the dependent and independent variables.

- Under regression analysis, listwise deletion is more robust than maximum likelihood and multiple imputation when MAR assumption is violated.

Listwise deletion (Note3)

- Disadvantages:
  - It will yield a larger standard errors than other more sophisticated methods discussed later.
  - If the data are not MCAR, but MAR, then your listwise deletion can yield biased estimates.
  - In other cases than regression analysis, other sophisticated methods can yield better estimates compared to listwise deletion.

Pairwise deletion (Notes1)

- This method could only be used in the case of linear models such as linear regression, factor analysis

- The premise of this method based on that the coefficient estimates are calculated based on the means, standard deviations, and correlation matrix

- As individual values will be impacted due to removal of some of the columns

- Compared to listwise deletion, we still utilized as many correlation between variables as possible to compute the correlation matrix.

Pairwise deletion (Notes1)

- Advantages:
  - If the true missing data mechanism is MCAR, pair wise deletion will yield consistent estimates, and unbiased in large samples
  - Compared to listwise deletion: ([Glasser 1964](#))
    - If the correlation among variables are low, pairwise deletion is more efficient estimates than listwise
    - If the correlations among variables are high, listwise deletion is more efficient than pairwise.

- Disadvantages:
  - If the data mechanism is MAR, pairwise deletion will yield biased estimates.
  - In small sample, sometimes covariance matrix might not be positive definite, which means coefficients estimates cannot be calculated.
  - **Note**: You need to read carefully on how your software specify the sample size because it will alter the standard errors.

Method 2: Imputation Methods

- Single Value Imputation
- Hot/Cold Deck Imputation
- KNN based Imputation
- Regression based Imputation
- Multiple imputation

Single value imputation

- **R**eplacing the missing value with a single value
  - Mean , Median , Most Frequent , Mean Person , of the corresponding feature .
- Dataset that have great outliers, Median is preferred

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean() →

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

- For encoded categorical features use Mode (Most frequent data value) within the column
- In some case use Zero, constant defined by user, Min, Max for the replacement

Single value imputation (Cont)

- **Pros:**
  - Easy and fast.
  - Works well with small numerical datasets.

- **Cons**:
  - Doesn't factor the correlations between features. It only works on the column level.
  - Not very accurate.
  - Doesn't account for the uncertainty in the imputations.

Hot or Cold Deck Imputation
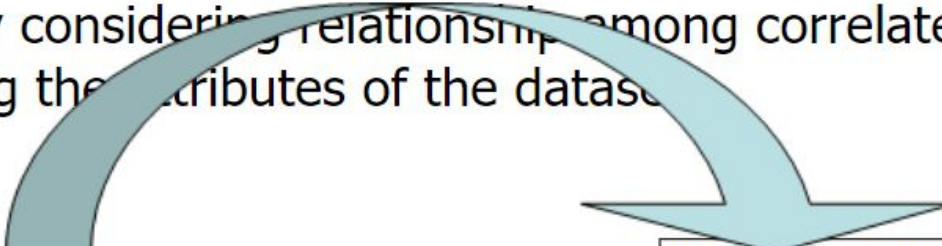
- **Hot Deck Imputation**
  - Find all the sample subjects similar on other variables, then *randomly* choose one of their values to fill in.
  - Good because constrained by pre-existing values
  - the randomness introduces hidden variability and is computationally expensive

- **Cold Deck Imputation**
  - Systematically choose the value from an individual who has similar values on other variables (e.g. the third item of each collection).
  - removes randomness of hot deck imputation.
  - Positively constrained by pre-existing values, but the randomness introduces hidden variability and is computationally expensive

Missing value replacement (Imputation)

- <u>Imputation</u> is a term that denotes a procedure that replaces the missing values in a dataset by some plausible values

  – i.e. by considering relationship among correlated values among the attributes of the dataset

Attribute 1	Attribute 2	Attribute 3	Attribute 4
20	cool	high	false
	cool	high	true
20	cool	high	true
20	mild	low	false
30	cool	normal	false
10	mild	high	true

Out of which one has missing value in Attribute1

If we consider only {attribute#2}, then value "cool" appears in 4 records.

Probability of Imputing value (20) = **66 %%**

Probability of Imputing value (30) = **33 %%**

Imputation (Example)

Attribute 1	Attribute 2	Attribute 3	Attribute 4
20	cool	high	false
	cool	high	true
20	cool	high	true
20	mild	low	false
30	cool	normal	false
10	mild	high	true

Imputation (Example)

Attribute 1	Attribute 2	Attribute 3	Attribute 4
20	cool	high	false
	cool	high	true
20	cool	high	true
20	mild	low	false
30	cool	normal	false
10	mild	high	true

For {attribute#4}   the value "true" appears in 3 records

Probability of Imputing value (20)   = 50%

Probability of Imputing value (10)   = 50%

Imputation (Example)

Attribute 1	Attribute 2	Attribute 3	Attribute 4
20	cool	high	false
	cool	**high**	**true**
20	cool	high	true
20	mild	low	false
30	cool	normal	false
10	mild	high	true

For {attribute#4} the value "true" appears in 3 records

Probability of Imputing value (20) = 50%

Probability of Imputing value (10) = 50%

Attribute 1	Attribute 2	Attribute 3	Attribute 4
20	cool	high	false
	cool	**high**	**true**
20	cool	high	true
20	mild	low	false
30	cool	normal	false
10	mild	high	true

Imputation (Example)

Attribute 1	Attribute 2	Attribute 3	Attribute 4
20	cool	high	false
	cool	**high**	**true**
20	cool	high	true
20	mild	low	false
30	cool	normal	false
10	mild	high	true

For {attribute#4} the value "true" appears in 3 records

Probability of Imputing value (20) = 50%

Probability of Imputing value (10) = 50%

Attribute 1	Attribute 2	Attribute 3	Attribute 4
20	cool	high	false
	cool	**high**	**true**
20	cool	high	true
20	mild	low	false
30	cool	normal	false
10	mild	high	true

For {attribute#2, attribute#3} the value {"cool", "high"} appears in only 2 records

Probability of Imputing value (20) = 100%

Imputation using KNN

- Algorithm for simple classification
- '**feature similarity**' to predict the values of any new data points
  - the new point is assigned a value based on how closely it resembles the points in the training set.
  - finding the *k's* closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood.
- **Pros:**
  - Can be much more accurate than the mean, median or most frequent imputation methods (It depends on the dataset).
- **Cons:**
  - Computationally expensive. KNN works by storing the whole training dataset in memory.
  - K-NN is quite sensitive to outliers in the data (**unlike SVM**)
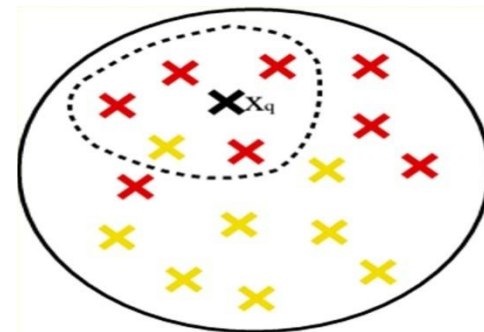
Single Regression based Imputation

- In single regression imputation the imputed value is predicted from a regression equation , we assume that the missing values are in a regression line with a nonzero slope with one of the complete features (predictors)
- Fill in with the predicted value obtained by regressing the missing variable on other variables; instead of just taking the mean, you're taking the *predicted value*, based on other variables.

Linear Regression

- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- $Y = \alpha + \beta X$
  - Two parameters , α and β specify the line and are to be estimated by using the data at hand.
  - using the least squares criterion to the known values of

$Y_1, Y_2, \ldots, X_1, X_2, \ldots$
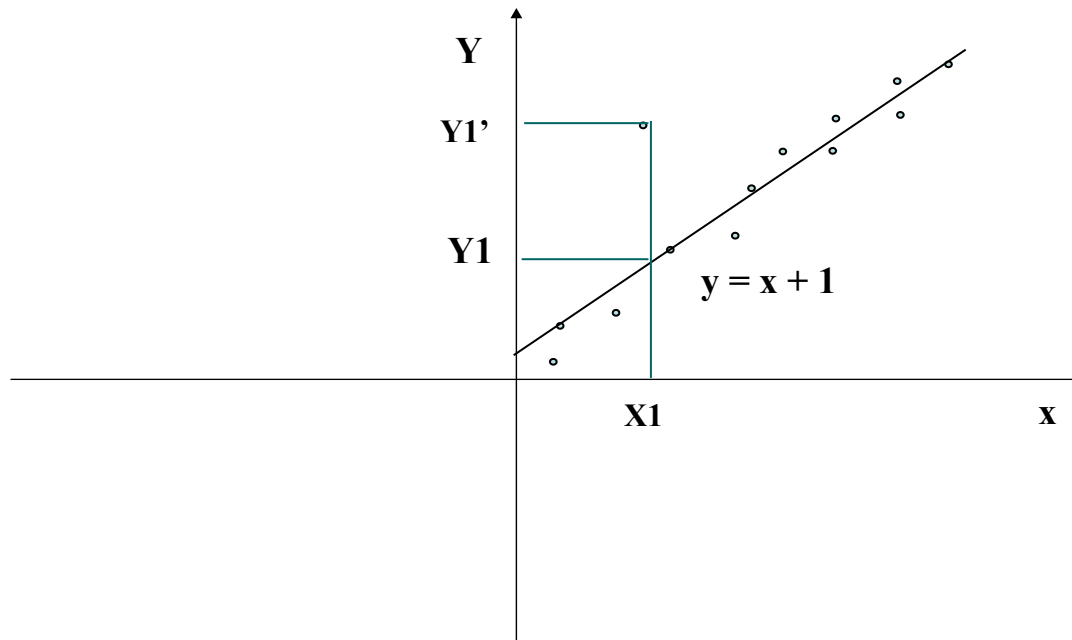
Linear regression based Imputation

- Disadvantage :
  - Random variation (i.e. an error term) around the Regression slop is not considered
  - Imputed values are therefore often too precise and lead to an overestimation of the correlation between X and Y

- Stochastic Regression

Random or stochastic regression imputation

- Appropriate random residual is added to the value predicted using regression mean imputation.

- Disadvantages
  - This method may lead to implausible or inappropreate values.
    - Variables are often restricted to a certain range of values (e.g. income should always be positive). Regression imputation is not able to impute according to such restrictions
  - This method leads to poor results when data is **heteroscedastic**
    - When data is skewed and demonstrate unequal variability inn the regression graph
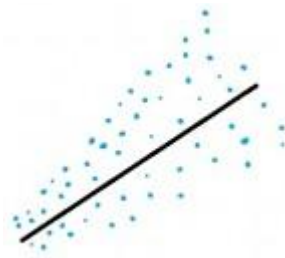
    - The imputation method assumes that the random error has on average the same size for all parts of the distribution,
    - often resulting in too small or too large random error terms for the imputed values.

Multiple regression based Imputation

- Multiple regression: allows a response variable Y to be modeled

  as a linear function of multidimensional feature vector

  - $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many nonlinear functions can be transformed to multiple regression

Log-Linear regression based Imputation

- Larger sample of observations more likely to demonstrate non-linear dependencies between dependent and independent variables (X,Y)
- Need to develop a non-linear model
  - require more advanced estimation techniques and computation power
  - Approximate non-linear relations with the mean of linear models on transformed variables
  - **The difference between the log-linear and linear model lies in the fact, that in the log-linear model the dependent variable is a product, instead of a sum, of independent variables**

$$Y_i = X_{1i}^{\beta_1} X_{2i}^{\beta_2} \ldots X_{ki}^{\beta_k} e^{\epsilon_i}$$

  - This model can be easily transformed into a linear model by taking a logarithm of each side of the above equation

Log-Linear regression based Imputation

- This model can be easily transformed into a linear model by taking a logarithm of each side of the above equation

$$Y_i = X_{1i}^{\beta_1} X_{2i}^{\beta_2} \dots X_{ki}^{\beta_k} e^{\epsilon_i}$$

$$log(Y_i) = \beta_1 log(X_{1i}) + \beta_2 log(X_{2i}) + \dots + \beta_k log(X_{ki}) + \epsilon_i$$

- Substitute
$$y_i = log(Y_i)$$
$$x_{ni} = log(X_{ni})$$

  where n = 1, 2, …, k, we obtain a purely linear model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

- Usage:
  - Due to log value can be used for only positive numeric values
  - when dependent and independent variables have lognormal distributions use log-linear regression
  - When those variables are normal or close to normal then use a simple linear model.
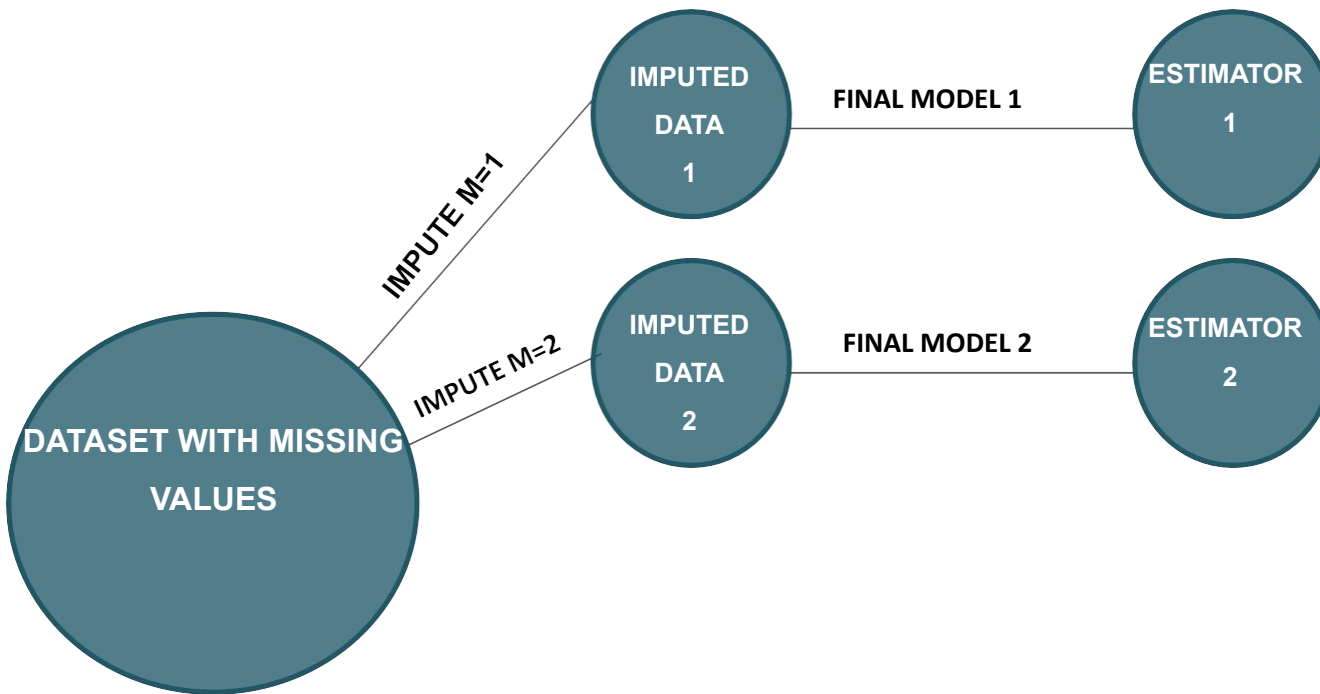
Multiple Imputation

- **Challenges of Single value imputation:**
  - values found in single imputation might be biased by the specific values in the current data set
  - not represent the total values of the full population

- **Sol : Multiple imputation**
  - Imputation techniques that assign several imputed values to each missing value using the following procedure
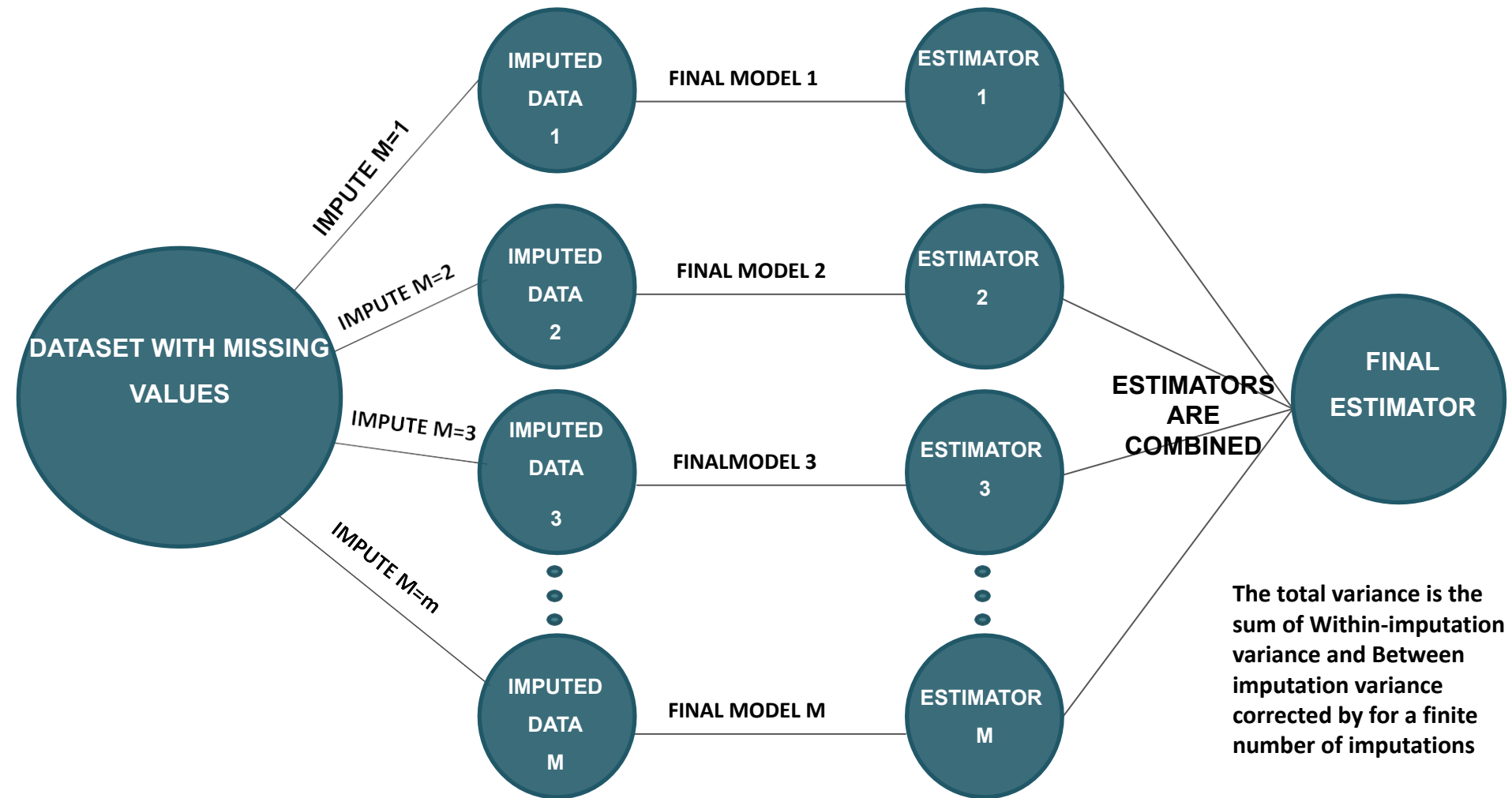
Multiple Imputation
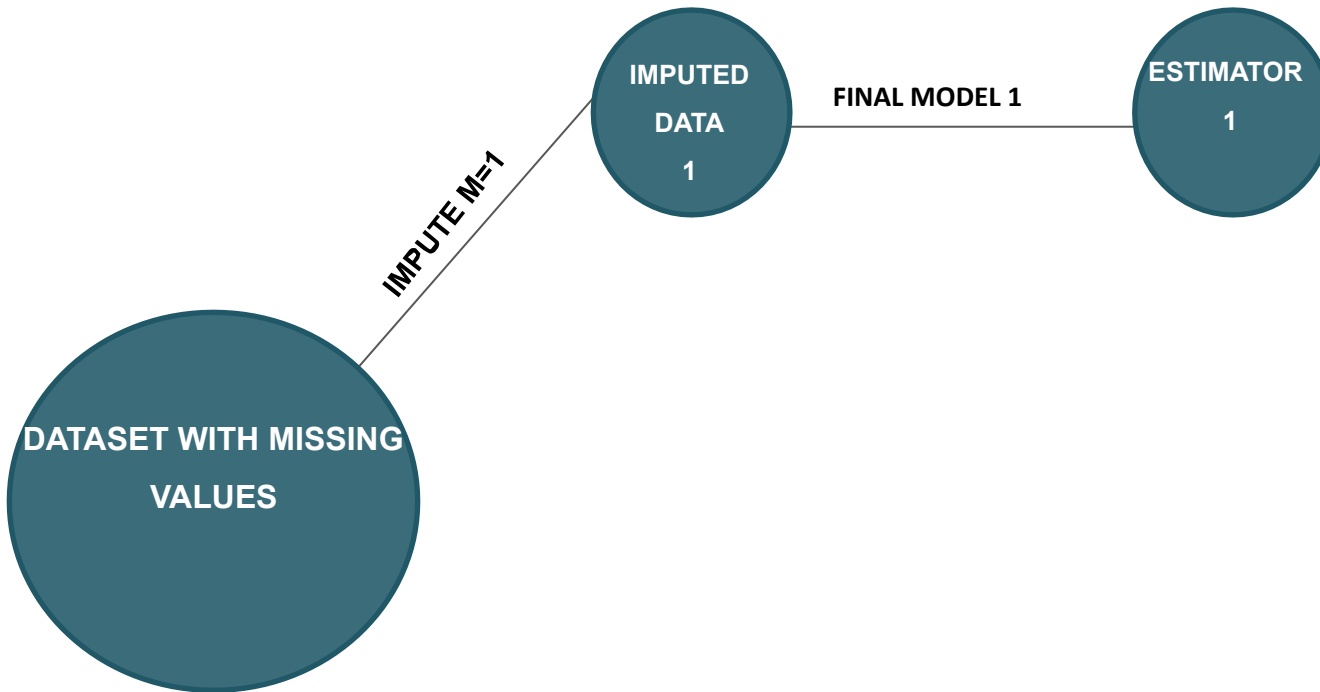
Multiple Imputation

Multiple Imputation

Multiple Imputation

Method:

- Multiple imputation method would run analyses with 5–10 unique samples of the dataset

- run the same predictive analysis on each**.

- Take mean of each result

- more numbers of repetitions will outcome less biased result

- Analyze spread of the result

- If they're clustering, they have a low standard deviation.
  - If they're not, variability is high and may be a sign that the value prediction may be less reliable.

- If this method is much more unbiased
  - it is also more complicated and requires more computational time and energy.