- The Ministry of Manpower (MOM)
- The Urban Redevelopment Authority (URA)
- Singapore Civil Defence Force (SCDF)
- The National Environment Agency (NEA)
- Building and Construction Authority (BCA)
- The Public Utilities Board (PUB)

Transform storey

- - to be 1
- Lx or Bx to be transformed
- G
- K, BK to be transformed (?) **currently left it as it is**
- Extracted level num from alphanumeric num

- 01/02 at alexandra rd. 2 storeys → hardcode?
- 02M at tuas. 629563,

Resi

- Top date. When the building is ready for use
- Explore ways to predict reconstruction date
- Reconstruction top date: when the building was tore down
- Postal_code6 transformation
- Top_date just want the year
- Storey_ht (storey height): get max level of building
- Got B1/B2. Transformed to -1, -2
- Transform postal codes to 6 char. But if postal codes are all integer then no need to be in strings
- Year of development. But should i look for it per building/per development (how to do so? Based on road name?)

Resi trans

- Caveat_date: when a transaction for housing takes place
- Nett price vs price
- Address
- Floor land area
- Title (e.g. freehold)

PUB

- Num of accounts
- hasDom
- hasNonPortable
- Do i extract storey_no max to get storey_ht? But simply to apply max need if else → increase complexity. Or use num_storey (ie. count num of levels)
- Then join with resi to get true storey_ht
- Some level is - some is NA

<u>Resi + PUB</u>

- Building height from max of Resi+PUB
- <u>Standardize column names, data values</u>
- <u>Standardized value types</u>
- Postal codes → check w Alvin again on code

<u>Measuring data accuracy</u>

- Weighted average?
- Ground truth?

- Measuring data quality:
    - https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf
    - Data completeness %
    - Data uniqueness % (e.g. address) → **number of things in real world**/number of things in data set
    - Validity → conforms to syntax (format, type, range)
    - Accuracy → count of accurate objects/(count of accurate + inaccurate objects)
    - Consistency %

- Ratio of data to errors: monitors the number of known data errors compared to the entire data set.
- Number of empty values: counts the times you have an empty field within a data set.
- Data time-to-value: evaluates how long it takes you to gain insights from a data set. There are other factors influencing it, yet quality is one of the main reason this time can increase.
- Data transformation error rate: this metric tracks how often a data transformation operation fails.
- Data storage costs: when your storage costs go up while the amount of data you use remains the same, or worse, decreases, it might mean that a significant part of the data stored has a quality to low to be used.

<u>Getting values</u>

- TOP missing values
    - able to get tax form for that building? (earliest)
- Look at power vs water utilities
- Occupancy
    - Able to get data related to trash collected?
- Lease data from owner vs whether there's any actual rental on govt db

- Or look at where a migrant worker is registered (to MOM) vs where he usually resides based on telco data?

<u>Week 3: Assessment of situation & solution & what is the action</u>

To do

- Standardization
    - Resi
        - Storey_no: -
        - Unit_no: -
        - storey_ht: NA
- Consistency
- Use central tendency to fill in missing vals

- Resi
    - 9 columns
    - 2,332,305 rows

    - Storey_no has weird values. Like ‘-’. Transformed to NA
        - Landed properties with multiple levels are treated as 1 storey ie. 1 unit = 1 storey
        - Transform_storey method to transform most cases. E.g. B1 = -1, L1 = 1, alphanumerical levels will have numeric levels extracted. B, K, BK etc are exceptions
        - Remove leading 0 from storey_no (number of distinct values still the same for storey_no and storey_no_new)
    - Postal_code6 transformed to be 6 char
        - Dynamic transformation - if original postal_code6 is 4 digit, append 2 ‘0’s in front
    - Year = max of top_date and reconstruction_date
        - ~8% of all values are NA
    - Year_of_development = max of year within the group (postal_code6, block_no)
        - Empty cells replaced with NAs, otherwise 2 blank cells will result in infinity
        - Standardize cells that have no values to be NAs
    - Storey_ht = max of storey_ht and storey_no_new
    - Num_storey = number of unique storeys in a building

    - Initial storey_ht = max of storey_ht and storey_no_new
        - **What happens if storey_no_new is alphabet/weird value?**
    - **So between storey_ht and num_storey, do we need to create a new column?**
    - 716,565 blocks (about 33%) with differing storey_ht and num_storey
        - Min -2

        - 1st quartile 1
        - Median 1
        - Mean 2.72
        - 3rd quartile 3
        - Max 40
    - 406,529 blocks with difference of 1 (most likely start counting num_storey from L2)

- PUB
    - Standardized to have same column names as Resi
    - Properties with storey_num as NA has all storey_ht as NA
    - Some blocks with storey_no and unit_no NA are actually L1 of the building
        - Num_storey transformation method picks it up as + 0 level
    - Transform_storey method to transform most cases. E.g. B1 = -1, L1 = 1, alphanumerical levels will have numeric levels extracted. B, K, BK etc are exceptions
        - Another exception is 01/02 at Alexandra Rd. hard code to get 02 as the storey_no
    - Postal_code6 transformed to be 6 char
        - Dynamic transformation - if original postal_code6 is 4 digit, append 2 ‘0’s in front
    - Num_storey = number of unique storeys in a building
    - New variables hasDomestic, hasNonPortable billing classes, num_accounts count number of unique accounts (0-2)
    - **Should i account for special level numbers? E.g. C2, G3 as irregularities**
    - Initially need to transform 3,772 records of storey_no
    - 5/1,593,779 records of storey_no left to further manual transform after initial transformation
    - **Choosing between storey_ht and num_storey**
    - Created storey_ht which is of storey_no and num_storey
    - 585,487 blocks (about 33%) with differing storey_ht and num_storey
        - Min 1
        - 1st quartile 1
        - Median 1
        - Mean 1.536
        - 3rd quartile 1
        - Max 57

- Ema
    - Storey_no extremely dirty
        - Transform. Values after / split into another data row
        - 2 same stories within a value sep by ‘/’. 64A/64B
        - 2 different stories within a value sep by ‘/’. 95/97
        - Sep by ‘ ’

        - Sep by ‘,’
        - Sep by ‘&’

        - Pseudo logic 1
            - Separate by [/,.&], if next char is just alphabet, take first number and append
            - Count number of units per row
            - Create new columns **storey_no_split** and **no_units_in_group**

        - **Pseudo logic 2**
            - Separate by [/,.&], if next char is just alphabet
            - Replace by ,
            - Separate column value by , and duplicate row
            - Faster than logic 1 cos lesser for loops

        - Cases
            - 50A/B → [50A, B] → 50A, 50B
            - 50/A/B → [50, A, B] → gives 50, 50A, 50B
            - 50A/B/60 → [50A, B, 60] → gives 50A, 50B, 60

            - If first number is just number, second number is alphabet

    - Filtered out ‘dirty’ values to **ema_irregular_vals** → requires human interpretation
    - ‘Cleaner’ values kept in **ema_regular_vals**

    - Storey_no
        - 1.7mil rows, filter away rows that are non-transformable e.g. already numeric, NA

    - SSIC
        - Has 0 values, NA
        - For units with domestic_ind=’DOMESTIC’, ssic is NA. otherwise will have ssic

- Resi_pub
    - Join by postal_code6, block_no, storey_no, unit_no

Resi_pub joins ema

- Should we standardize all empty values to NA? Example block and unit no NA in one table, empty in the other. Thus, didn’t join
- What does it mean when storey_no/unit_no is ‘-’?

- Storey_ht has NA/0

<u>Resi_pub</u>

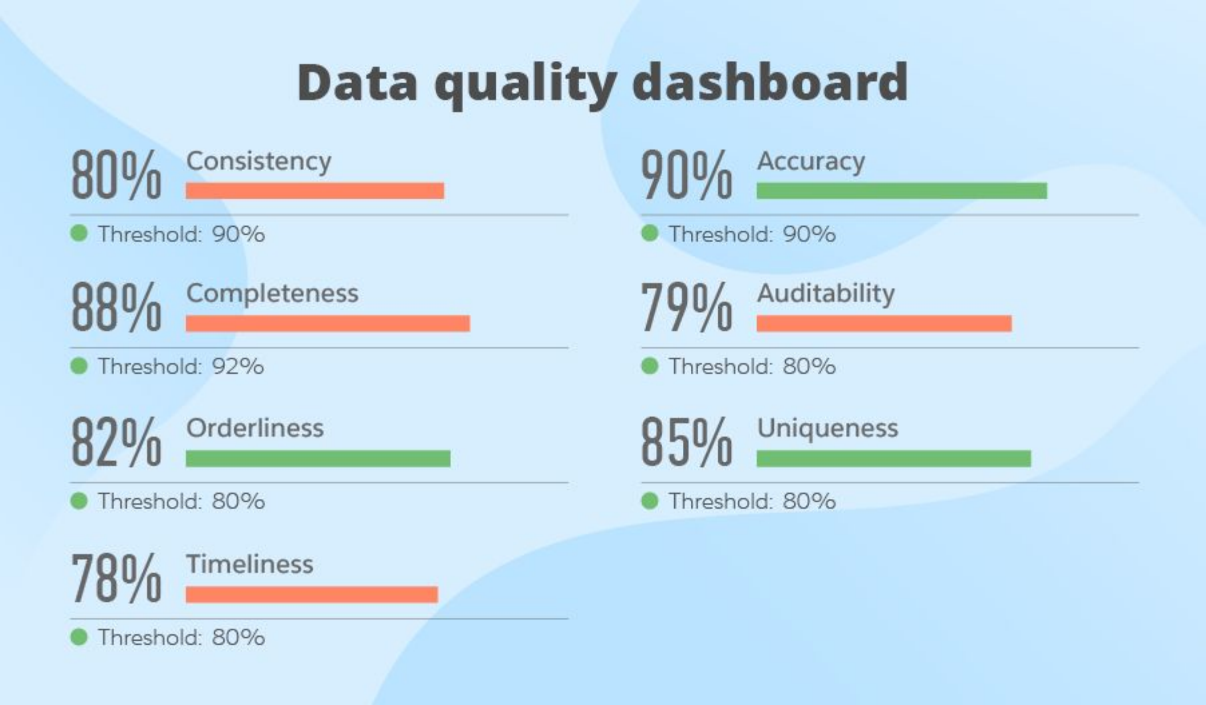
- Remove storey_no, postal_code6 as they have been transformed
- Compare across storey_ht and num_storey and get the maximum value across
    - Consideration: num_storey looks at number of unique storey_no values, storey_ht looks at number of storeys from G. may be slightly inaccurate?
        - Maybe num_storey's calculation exclude <1?
- Keep either Resi or PUB's road_name value
    - Currently Resi takes precedence
- Resi outer_join pub
    - Resi: 2,332,305
    - PUB: 1,593,779
        - Sum: 3,926,084
    - Resi_PUB: 3,550,781
        - No common value between Resi and PUB: 1,218,476 **[HDB units]**
    - **Is_resi: is it more useful to label 1, 2, 3 (resi, pub, ema)???**

<u>Resi_pub_ema</u>

- Single postal code can belong to multiple blocks. E.g. 640258 belonging to blk 515, 509 etc in **EMA**
- Remove postal_code6, storey_no, storey_no_split
- Street_name vs road_name
- Didn't include **ema_irregular_val**

<u>Data quality</u>

- Rank importance of features
- Completeness
- Validity
    - PUB weird values in storey_no
- Data set quality score = average of data quality score for all columns

- 

Data quality dashboard

- Able to think about data rules? Standardizing future data collection?

<u>Predicting year of development</u>

UK:
https://www.researchgate.net/publication/326920098_Predicting_residential_building_age_from_map_data

- Area of building
- Perimeter of building
- Building height
- Type of housing e.g. detached, semi-d

- 40% of missing values in the resi_pub_ema
- ~8% missing in resi
- 1,408 lesser NAs after property guru data was imputed
- Still ~40% missing values

- Completeness
- Accuracy
-

- Identify similar units
- Impute missing values based on dataset
- KNN?

Values that should exist:

- Postal code
- Road name

- Complete case analysis
    - Use only when missing data <5%
    - Drop rows with missing values
- Mean, median, mode
    - Not accurate
- Regression imputation?
    - Check relationships between variables
- KNN imputation
    - Look for k nearest neighbours, based on mean/median
- Multivariate imputations
    - Cannot be used as the variables might be correlated

- Unit count
- 139A/B/C maybe from shophouses
- Spatial data and building type
- Age of building information
- Storey height vs storey number. tbc
- Skeleton of slide
- Penalties for different imputation methods
- Irregular storey number is humanly possible to edit
- Joined table vs IRAS
- Spatial patterns

PUB

- Storey_no to be transformed: 3772 out of 1,593,779
- Orchard Rd 238842 with weird storey_ht values
    - K, BK, NA, NA
    - Use case for num_storey for such situations
    - > 75% of difference in num_storey and storey_ht is 1
        - Could be L1

EMA

- Storey_no with weird values turn out to be building/blk no
- 91 out of 1,745,890 values require manual intervention

Ultimate table

- ~30% of data where storey_ht does not equate to num_storey
- 4,154,914 rows
- ~39% missing values in TOP date
- @ PUB accounts
    - Only 2 3-Room HDB units have 2 accounts

    - ⅔ of the dataset are private apartments with 1 account. Domestic usage

- KNN on current dataset
    - Mix of qualitative and quantitative attributes
    - Need to transform address related data into geospatial data
    - Airbnb, but all features are quantitative: https://www.dataquest.io/blog/machine-learning-tutorial/

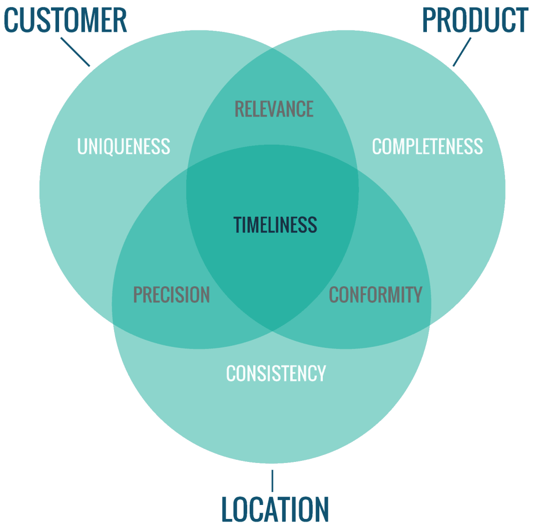
1. Encode the categorical variables
2. Join resi_pub_ema_test_sz to pub_update for water usage
3. Get xy coordinates of postal codes
4. Run KNN based on postal codes, premise type, subsector, domestic_ind, **median water usage**
    a. KNN classification for categorical outcome
    b. KNN regression for continuous outcome
5. Investigate effect of geospatial location on KNN
6. Consideration of using purely clean rows to do KNN (impossible, only 990 rows are clean)
    a. Consider dropping non important columns and calculate again
7. KNN on current dataset
    a. Maybe no need ref period since pub water data would be sufficient?
    b. TOP_date
    c. Is_resi=1
    d. hasDomestic
    e. hasPortable
    f. Num_accounts
    g. Storeyht_numstorey_max
    h. Domestic_ind
    i. Sub_sector
    j. premises_type

- Encode:
    - Domestic_ind
    - Premises_type
    - sub_sector

- Join enforcement table to get
    - Median and mean of electricity consumption
    - Nov 2017 to Aug 2019

- Left with:
    - KNN
    - KNN by geospatial data **(not possible as onemap is not avail offline)**
    - Validate imputation quality score

- Analysis of enforcement data
    - All happened in private resi
    - The only type of HDB involved is 5-room. 1 unit only
    - Thus look at resi data
    - Non-resi data's private data is minimal
    - Compare data of water usage normal resi vs enforcement
    - hotspots : geylang, bedok, kallang, rochor
    - 878 out of 1.7mil rows are outliers (>10000 in water consumption)

1. Rows with missing values in multiple columns
2. Data imputation & measuring the accuracy of imputation
    a. Data integration from other sources e.g. Property Guru (https://www.researchgate.net/publication/281284460_Big_Data_Pre-Processing_A_Quality_Framework)
    b. https://www.hindawi.com/journals/mpe/2015/538613/tab3/
3. Data Quality Quantification
    a. 

    b. https://profisee.com/data-quality-what-why-how-who/
4. Identify similar units (but with missing values)
    a. Complete case analysis followed by partial clustering https://www.displayr.com/5-ways-deal-missing-data-cluster-analysis/

https://arxiv.org/ftp/arxiv/papers/1603/1603.03281.pdf

1. Join resi, pub, ema (left join all the way)
2. Remove unit_no and check number of complete rows
3. Cluster rows with no missing records
4. Find distance of each row in missing records table to cluster centre

- Will zoning details even be useful?
- About 62% of data is complete (before taking out storey and unit num)
- 78% complete data after taking out storey and unit num

- 6 times jump in enforcement data?! Pls investigate
- No labels for the individual units?

3 types of missing data:

1. MCAR

- Probability of missingness completely unrelated to any variables in dataset
- Complete case approach

2. MAR

- May be systematic missingness in relation to another variable being measured/probability of missingness depends on other variables in dataset

3. MNAR

- Certain patterns in missingness/missingness related to own variable vale
    - Selection models
        - Specify joint distribution through the marginal distribution P(Y) of the measurements and conditional distribution (R|Y) of missing data given the measurements
        - Require strong assumptions to describe the potential dropout patterns
        - Factors joint distribution into the marginal measurement model that describes the distribution of complete measurements, and the dropout model describes the conditional distribution of dropout indicators, given the observed and unobserved measurements
        - **Conclusions obtained depend on assumptions made some of which cannot be investigated from the data under analysis**
    - Pattern mixture models
        - Variables of interest are assumed to have distributions according to whether they are missing/non-missing
        - Specify joint distribution through marginal distribution of missing data and conditional distribution of measurements given missing data
        - Models are under-identified; that is, for each dropout pattern the observed data does not provide direct information to identify the distributions for the incomplete patterns
        - Resolves under-identification problem through the use of identifying restrictions
    -
    - ML/MI (MAR analysis models for MNAR data are still pretty good)

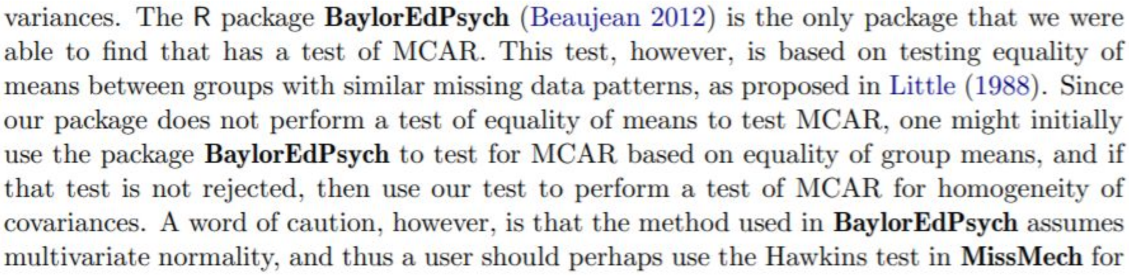
- Selection models
    - Let Y be the variable that user wants to specify
    - Specify the marginal distribution P(Y) and response weights P(R|Y). P(Y) and P(R|Y) are unknown, to be specified by user
    - P(Y) normally distributed
    - Correct for selection bias

    - Distribution should be a realistic description of combined observed and missing values
- Pattern mixture models
    - Reverse the role of Y and R
    - Emphasizes that combined distribution is a mix of distributions of Y in responders and nonresponders
    - Can estimate difference in mean of observed vs missing values
    - Assumes that the variables of interest have distributions according to the status missing or non-missing

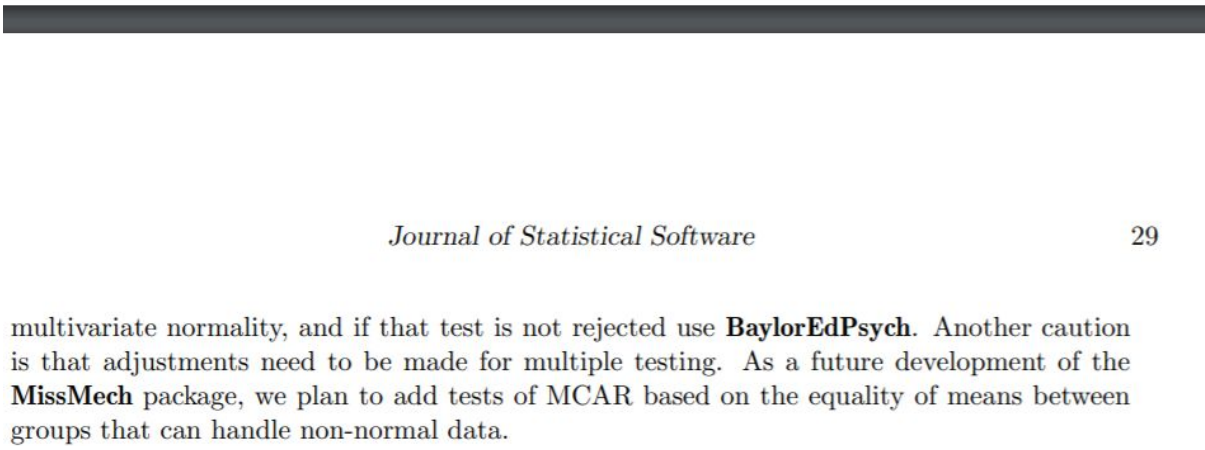
Determine missingness

1. Shadow matrix
    - Indicator variables where 1 is present and 0 is missing
        - Variables tend to be missing together = MAR
        - Variables not missing together = MCAR
2. Test missingness for independence
    - http://www.statisticalassociates.com/missingvaluesanalysis_p.pdf
    - Binary variable where 0 = not missing, 1 = missing
    - Statistical test to see if other modeled variables are correlated with binary missingness indicator variable with each measured variable
    - Checks MCAR vs MAR
3. Little's MCAR test
    - http://www.i-deel.org/uploads/5/2/4/1/52416001/chapter_4.pdf
    - https://www.academia.edu/2034404/A_Probability_Based_Test_for_Missing_Completely_at_Random_Data_Patterns_a.k.a_Extending_Littles_Missing_Completely_at_Random_Test_
    - Tests whether significant difference exists between means of different missing-value patterns
    - Null hypothesis is that the data is MCAR. If p-value < 0.05, reject the null hypothesis that data is MCAR. Otherwise there is not enough evidence to reject = possible that data is missing randomly
    - **BaylorEdPsych** R Package
        - Test based on means under different missing data patterns
        - Test statistic is a weighted sum of standardized differences between subgroup means and grand means
        - But assumes multivariate normal distribution https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3124223/
        - BaylorEdPsych test equality of means between groups with similar missing data patterns
    - MissMech (can work with non-parametric), test homogeneity of covariances file:///C:/Users/d-ebb/Downloads/v56i06%20(1).pdf
        - But assumptions do not support categorical variables

- **Word of caution:**



variances. The R package **BaylorEdPsych** (Beaujean 2012) is the only package that we were able to find that has a test of MCAR. This test, however, is based on testing equality of means between groups with similar missing data patterns, as proposed in Little (1988). Since our package does not perform a test of equality of means to test MCAR, one might initially use the package **BaylorEdPsych** to test for MCAR based on equality of group means, and if that test is not rejected, then use our test to perform a test of MCAR for homogeneity of covariances. A word of caution, however, is that the method used in **BaylorEdPsych** assumes multivariate normality, and thus a user should perhaps use the Hawkins test in **MissMech** for



multivariate normality, and if that test is not rejected use **BaylorEdPsych**. Another caution is that adjustments need to be made for multiple testing. As a future development of the **MissMech** package, we plan to add tests of MCAR based on the equality of means between groups that can handle non-normal data.

-
- & because it replicates so many versions of same dataset, RAM will run out. So another alternative is to take a random sample of 10,000 rows out to run the MCAR test

4. Expert indication

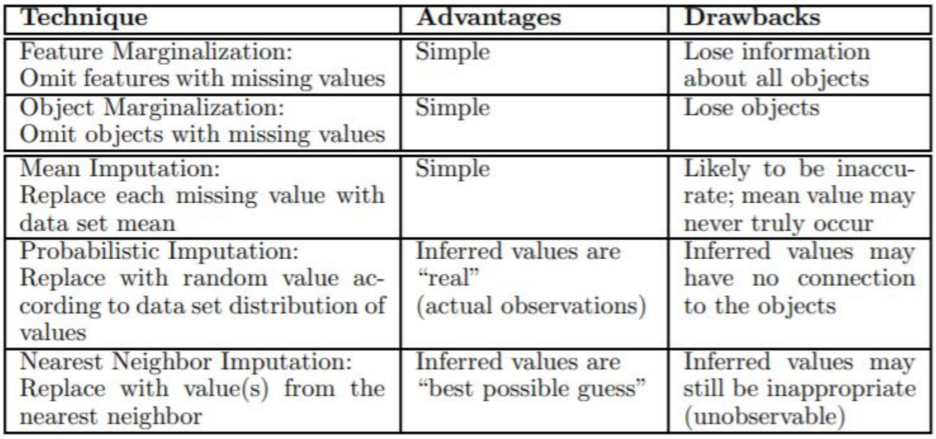
- Almost impossible to prove MNAR (MAR VS MNAR)
- Can test for MCAR

Points to consider

- Factors associated with having a complete record

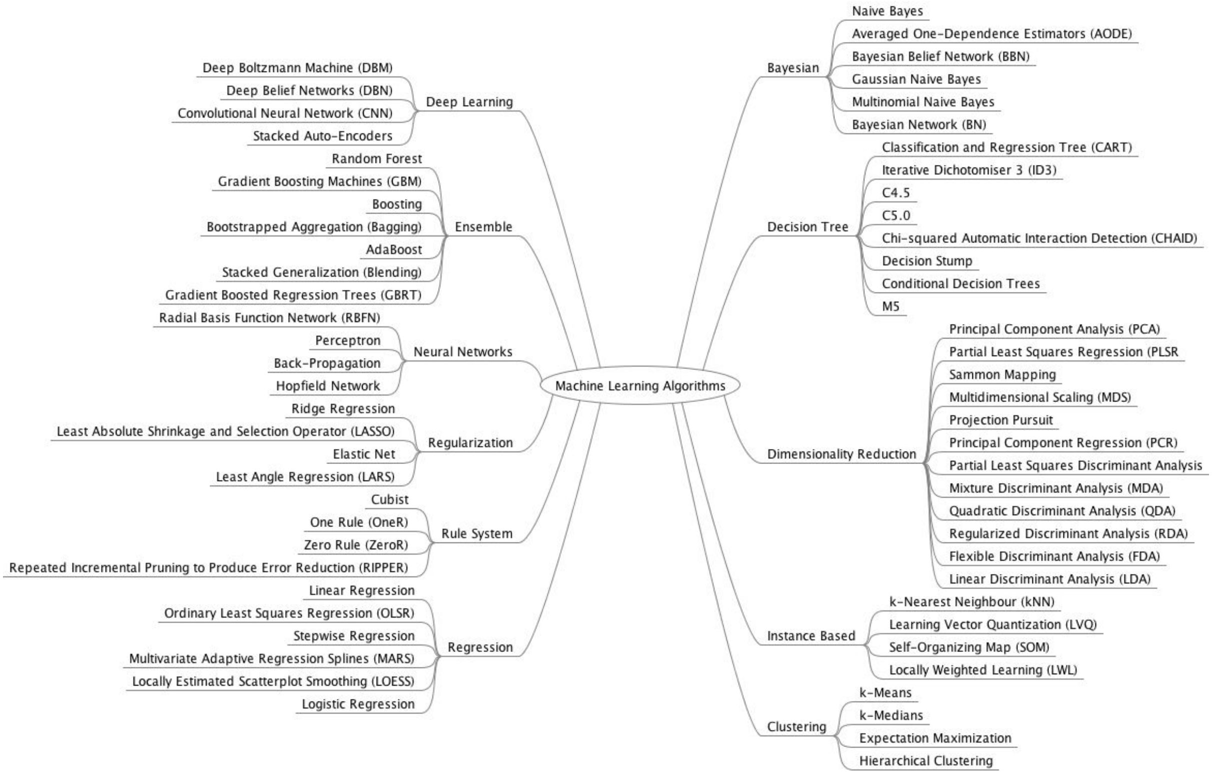
Lit review

- **Filling in missing data + accuracy (without labels/ground truth)**
    - Techniques: chosen data imputation



Technique	Advantages	Drawbacks
Feature Marginalization: Omit features with missing values	Simple	Lose information about all objects
Object Marginalization: Omit objects with missing values	Simple	Lose objects
Mean Imputation: Replace each missing value with data set mean	Simple	Likely to be inaccurate; mean value may never truly occur
Probabilistic Imputation: Replace with random value according to data set distribution of values	Inferred values are "real" (actual observations)	Inferred values may have no connection to the objects
Nearest Neighbor Imputation: Replace with value(s) from the nearest neighbor	Inferred values are "best possible guess"	Inferred values may still be inappropriate (unobservable)

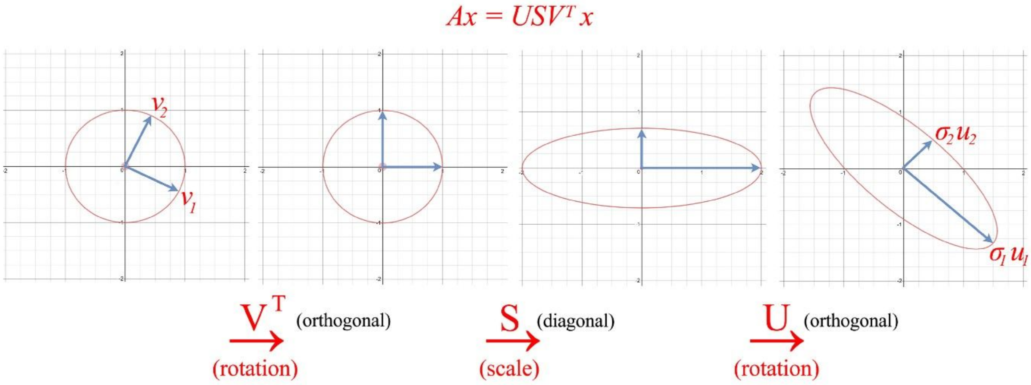
- Imputations always happen on the assumption that missing data occurs by random chance
- Can do data deletion but that would skew the dataset
  - Recommended for dataset with missing observations of <5%
- IMI: https://www.researchgate.net/publication/286869857_Classification_Uncertainty_of_Multiple_Imputed_Data
  - Split into training and test
  - Training generate multiple datasets using MICE
  - Stack all training imputed sets together and use it to train classifier
  - Trained classifier use on test set


Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

https://towardsdatascience.com/which-machine-learning-model-to-use-db5fdf37f3dd

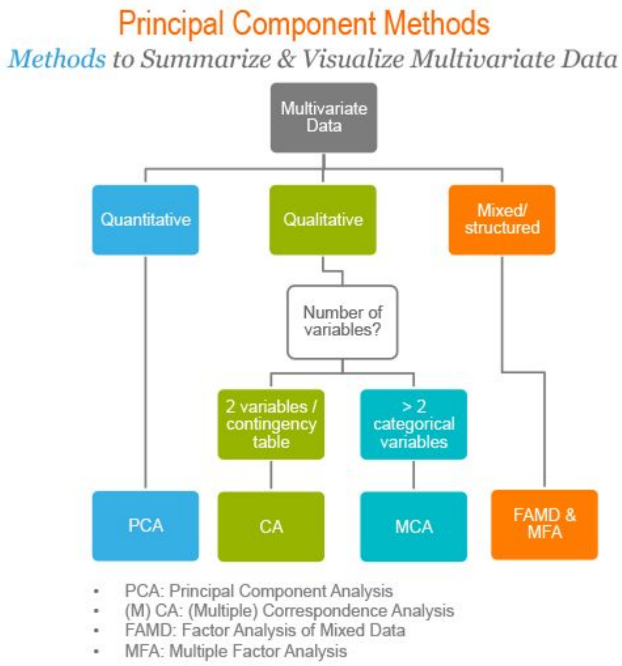
Imputation techniques

- Main consideration against traditional models for continuous variables is that they require categorical variables to be encoded. But we cannot just simply assume encoded categorical variables to follow the assumptions made for continuous variables e.g. distribution patterns

1. **SVD impute**
    - https://arxiv.org/pdf/1804.11087.pdf
    - https://gregorygundersen.com/blog/2018/12/10/svd/
    - Matrix factorizations
        - https://gregorygundersen.com/blog/2018/10/24/matrices/
    - Dimension reduction: imputes data with principal component methods that take into account similarities between observations and relationship between variables
    - Gets rid of redundant data (eigenvalues in SVD help determine what variables are most informative, and what you can do without)
    - SVD: rotate axes of feature space in which a dataset has been plotted so instead of different axes corresponding to different features, new axes point in directions that track linear combinations of original feature space
        - New axes are useful because they systematically break down the variance in the data points based on each direction's contribution to variance in the data
        - Results of this process is a ranked list of directions in the feature space ordered from most variance to least
        - Directions along which there is greatest variance are "principal components" of variation in data
    - 

What is the difference between SVD and PCA? SVD gives you the whole nine-yard of diagonalizing a matrix into special matrices that are easy to manipulate and to analyze. It lay down the foundation to untangle data into independent components. PCA skips less significant components. Obviously, we can use SVD to find PCA by truncating the less important basis vectors in the original SVD matrix.
https://medium.com/@jonathan_hui/machine-learning-singular-value-decomposition-svd-principal-component-analysis-pca-1d45e885e491

2. **Factorial Analysis for Mixed Variables (principal component method)**


Principal Component Methods
Methods to Summarize & Visualize Multivariate Data
Multivariate Data
Quantitative
Qualitative
Mixed/ structured
Number of variables?
2 variables / contingency table
> 2 categorical variables
PCA
CA
MCA
FAMD & MFA
PCA: Principal Component Analysis
(M) CA: (Multiple) Correspondence Analysis
FAMD: Factor Analysis of Mixed Data
MFA: Multiple Factor Analysis

- http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/
- Principal component method for factorial analysis for mixed data (FAMD)
- Represent variables of a dataset linearly related to random unobservable variables called factors, aka latent variables. Their existence is hypothetical as they cannot be observed or measured
- Similar to PCA as it reduces original variables to smaller number of factors to allow for easier interpretation
- But different in:
  - Principal components are linear combinations of the variables
  - Factors are linear combinations of underlying latent variables
  - PCA explains total variance, FAMD accounts for covariances of correlations among the variables
  - Factor does not belong to variable's space. Not a function of variables
- Assumptions
  - Latent variables are independent of each other and of error terms
  - Covariance of error terms and factor is 0

- Initialization step - missing values imputed with initial values such as mean of variable for continuous values and proportion of category for each category using the non-missing entries (pg 5 on how categorical initial impute is done: https://arxiv.org/pdf/1301.4797.pdf)
- Second step is to perform FAMD on completed dataset
- Imputes missing values with reconstruction formulae of order ncp
- In short, principal components obtained from the initial 'complete' data is used to reconstruct the data and impute missing values
- Considers relationships between continuous and categorical variables

3. **Decision Trees and Decision Tree based ensembles like Random Forests and Gradient Trees**
    - **Random Forests**
        - Extension of CART, do not rely on distributional assumptions and can accomodate non linear relations and interactions
        - missForest in R. Random forest based algorithm for missing data imputation, can be used for mixed data type
        - Random forest = build multiple trees. To classify a new object based on attributes, each tree gives a classification and use voting to determine the type of classification. Takes average of outputs by different trees
            1. Assume number of cases in training set is N. a sample of N these cases is taken at random but with replacement
            2. M input variables/features, a number m<M is specified such that at each node, m variables are selected at random out of the M. Best split on these m is used to split the node. Value of m is held constant while forest is grown
            3. Each tree is grown to largest extent possible and there is no pruning
            4. Predict new data by aggregating predictions of the n trees (majority votes for classification, average for regression)
        - Iterative imputation technique based on Breiman’s Random Forests: iterative in that variables with least number of missing values are imputed first using predictions from the random forests. Repeated for each variable until predictions stabilize
        - A random forest trained based on observed values of a data matrix are used to predict the missing values
        - https://www.researchgate.net/publication/331092508_Comparison_of_Selected_Multiple_Imputation_Methods_for_Continuous_Variables-Preliminary_Simulation_Study_Results
        - Has better performance than KNN
        - Not available for imputation in R
        - https://arxiv.org/pdf/1701.05305.pdf
        - Estimates OOB error
            1. They don’t appear in all trees.
            2. Each time you train a tree, a fraction of the data is left out. This is the OOB or “out of bag” data from which the OOB error derives its name.
            3. You predict on that fraction each time.
            4. At the end, you’ve got a bunch of predictions on each sample from the trees it WASN’T in.
            5. Average those, and you have an OOB prediction, with which you calculate the OOB error.

MICE Random Forests

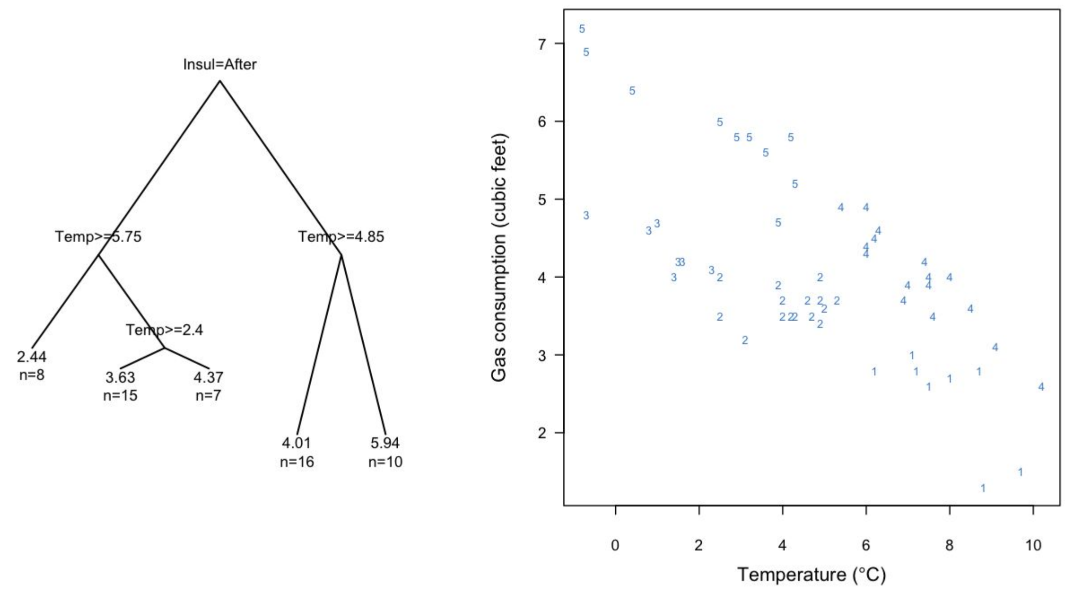
1. Form data matrix Y using complete dataset, and another data matrix Y initial

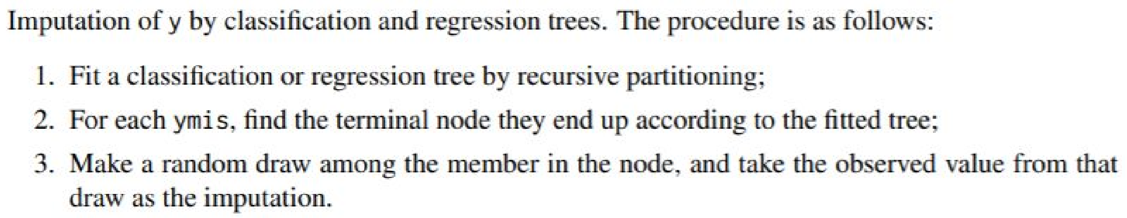
2. For the first variable w/ missing values, impute the missing observations with random draws from their same variable and update Y
3. Random forests first draws k bootstrap samples from complete portion of Y initial (Y complete initial)
4. One tree fitted for every bootstrap sample. Result is k trees, where each tree has several leaves. Each leaf includes a subset of Y complete initial which are called donors
5. For missing observations, determine which leaf they will end up according to the k trees
6. Take all donors from the k leaves and randomly select one observed value from the donors, then update Y
7. Repeat steps 2 to 6 to perform it x iterations
8. Repeat steps 1 to 7 m times to obtain m datasets

    - **CART**
        - Classification and regression trees
        - Non-linear relationship
        - Used in missForest. But mice provides multiple imputation

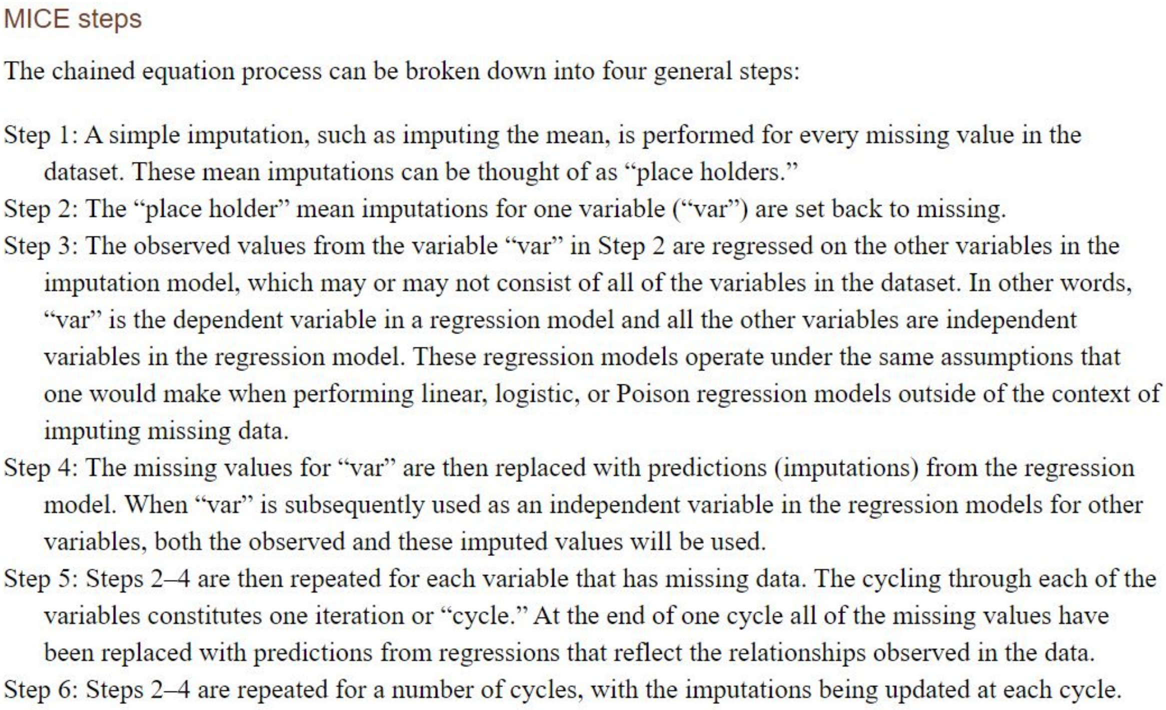
<u>MICE CART</u>

1. Build CART by recursive partitioning using complete observations to form data matrix Y
2. For the first variable w/ missing values, impute the missing observations with random draws from their same variable and update Y
3. For each missing value, find the terminal node they end up in
4. Make a random draw among the members in the node and the observed value will be the imputed value
5. Update imputed missing values to Y
6. Re-run steps 2 to 5
7. Repeat steps 1 to 6 m times to form m sets




Imputation of y by classification and regression trees. The procedure is as follows:

1. Fit a classification or regression tree by recursive partitioning;
2. For each ymis, find the terminal node they end up according to the fitted tree;
3. Make a random draw among the member in the node, and take the observed value from that draw as the imputation.


MICE steps

The chained equation process can be broken down into four general steps:

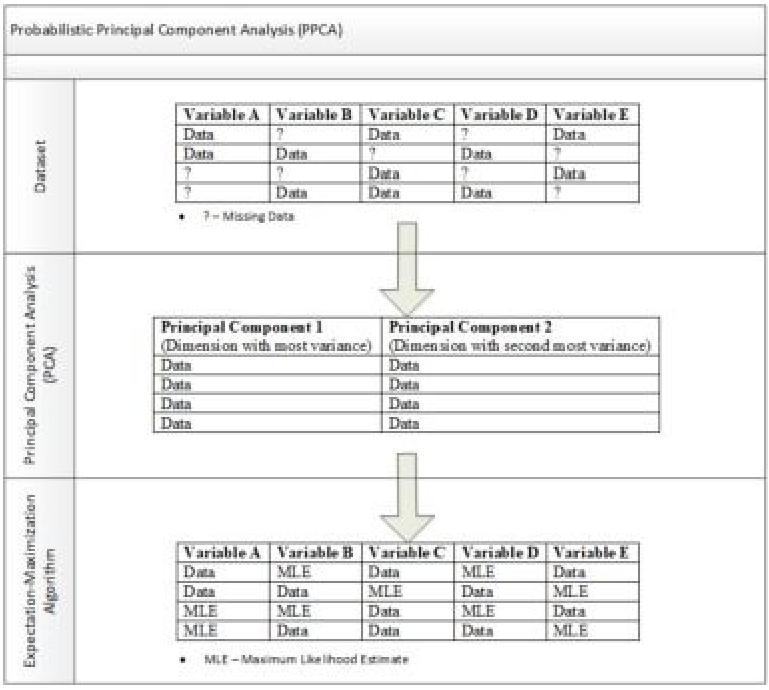
Step 1: A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
Step 2: The "place holder" mean imputations for one variable ("var") are set back to missing.
Step 3: The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model, which may or may not consist of all of the variables in the dataset. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model. These regression models operate under the same assumptions that one would make when performing linear, logistic, or Poison regression models outside of the context of imputing missing data.
Step 4: The missing values for "var" are then replaced with predictions (imputations) from the regression model. When "var" is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.
Step 5: Steps 2–4 are then repeated for each variable that has missing data. The cycling through each of the variables constitutes one iteration or "cycle." At the end of one cycle all of the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data.
Step 6: Steps 2–4 are repeated for a number of cycles, with the imputations being updated at each cycle.

- **MICE Random Forest**
    - https://www.researchgate.net/publication/260485157_Comparison_of_random_forest_and_parametric_imputation_models_for_imputing_missing_data_using_MICE_A_CALIBER_study

        - Records with missing values in dependent variable are imputed by random draws from independent normal distributions centred on conditional means predicted using random forest
        - Used out of bag mean squared error as an estimator of residual variance
        - Random forest fits each tree to a different bootstrap sample of the data and aggregates the results; OOB error is mean of squared differences between each observed value and prediction based on trees for which that observation is not included in the bootstrap sample

4. **Naive techniques: K-Nearest Neighbor Classifiers, Naive Bayes Classifier etc**
    - **KNN**
        - A type of hot deck method
        - Non-linear relationship
        - Sequential KNN
            1. Impute dimensions with fewest missing values and imputes missing values from previously imputed values
        - Iterative
            1. Iterative process to refine estimates and choose nearest neighbors based on estimates from previous iteration
        - Identifies k closest observations based on Gower distance (for mixed data) and computes weighted average of these k observation
        - Distance between 2 observations is weighted mean of contributions of each variable, where weight should represent the importance of the variable
        - Continuous variables imputed with median, categorical variables use mode (if tie then randomly drawn)
        - file:///C:/Users/d-ebb/Downloads/Imputation_with_the_R_package_VIM.pdf
    - **Naive Bayes**
        - Assumes variables are **conditionally independent** of one another & **Gaussian distribution** (e1071 package in R)
        - Assumes features contribute independently to the probability
        - More suited to categorical variables
        - But certain studies show that its possible to ignore the independence rule and get good imputation results e.g. https://rviews.rstudio.com/2016/11/02/naive-bayes-a-generative-model-and-big-data-classifier/
    -

5. Bayesian PCA



Probabilistic Principal Component Analysis (PPCA)

Dataset
Principal Component Analysis (PCA)
Expectation-Maximization Algorithm

Variable A	Variable B	Variable C	Variable D	Variable E
Data	?	Data	?	Data
Data	Data	?	Data	?
?	?	Data	?	Data
?	Data	Data	Data	?

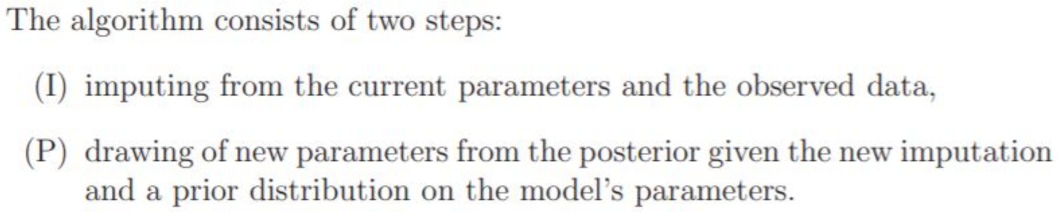
- ? – Missing Data

Principal Component 1 (Dimension with most variance)	Principal Component 2 (Dimension with second most variance)
Data	Data
Data	Data
Data	Data
Data	Data

Variable A	Variable B	Variable C	Variable D	Variable E
Data	MLE	Data	MLE	Data
Data	Data	MLE	Data	MLE
MLE	MLE	Data	MLE	Data
MLE	Data	Data	Data	MLE

- MLE – Maximum Likelihood Estimate

    - **https://papers.nips.cc/paper/1549-bayesian-pca.pdf**
    - BUT NOT BUILT FOR MIXED DATA
    - Automatic selection of appropriate model dimensionality
    - Conventional PCA does not define probability distribution
    - Bayesian needs to choose prior distribution and formulate tractable algorithm
    - Reflects uncertainty of the parameters from one imputation to the next
    - “Bayesian” is based on an iterative algorithm which alternates imputation of the data set and draw of the PCA parameters in a posterior distribution
    - Bayesian methods are very elegant, but require a shift in mindset: we are no longer looking for a point estimate of the parameters (as in maximum likelihood or MAP), but for a full posterior distribution.
    - Determine posterior distribution for model’s parameters using prior distribution and the observed entries
        - http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.753.8914&rep=rep1&type=pdf



The algorithm consists of two steps:

(I) imputing from the current parameters and the observed data,

(P) drawing of new parameters from the posterior given the new imputation and a prior distribution on the model’s parameters.

    -

Bayesian statistics steps:

1. Define the prior distribution that incorporates your subjective beliefs about a parameter (in your example the parameter of interest is the proportion of left-handers). The prior can be "uninformative" or "informative" (but there is no prior that has no information, see the discussion here).

2. Gather data.
3. Update your prior distribution with the data using Bayes' theorem to obtain a posterior distribution. The posterior distribution is a probability distribution that represents your updated beliefs about the parameter after having seen the data.
4. Analyze the posterior distribution and summarize it (mean, median, sd, quantiles, ...).

The basis of all bayesian statistics is Bayes' theorem, which is:
Posterior ∝ prior × likelihood

- Multiple imputation
    - **MICE imputation**
        - Explanation on how it works
            - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
            - Sequential regression multiple imputation
        - https://uvastatlab.github.io/2019/05/01/getting-started-with-multiple-imputation-in-r/
        - Chained equation because each variable can be modelled to its distribution
        - Assumes parametric distribution wa**so no?**
        - Joint multivariate normal distribution multiple imputation
            - Amelia/norm packages
            - **Does not perform well where there is no linear relationship between variables**
        - Conditional Multiple Imputation
            - Assumes distribution for each variable rather than entire dataset
            - But! Imputation output can only fit linear model
        - E.g. MICE/pmm imputation. R package of MICE allows for imputation of categorical variables **only for MAR**
            - **MICE statistical estimates?**
            - MICE vs mean vs deletion: https://www.researchgate.net/publication/262036960_Missing_Data_The_Importance_and_Impact_of_Missing_Data_from_Clinical_Research
            - Predictive mean matching
                - For each missing entry, a small set of candidate donors is formed from complete cases
                - One donor is randomly drawn from the set
                - Assumes distribution of missing cell is same as observed data of candidate donors
        - With CART/RF
            - https://stefvanbuuren.name/fimd/sec-cart.html
        - Number of datasets to be created depends on % of missingness
            - E.g. 10 sets if 10% missing http://www.emgo.nl/kc/handling-missing-data/

- **Amelia**
    - Assumption that dataset is multivariate normal
- **Hmisc**
    - http://finzi.psych.upenn.edu/R/library/Hmisc/html/aregImpute.html
    - aregImpute vs transcan
        - Use pmm
        - aregImpute takes into account of uncertainty in imputations
        - Restricted cubic splines: a way of testing hypothesis that relationship is not linear
        - https://towardsdatascience.com/restricted-cubic-splines-c0617b07b9a5
        - Number of knots (nk): recommended by Frank Harrell use 5 knots if is large dataset

Hmisc steps

In bootstrapping, different bootstrap resamples are used for each of multiple imputations. Then, a flexible additive model (non parametric regression method) is fitted on samples taken with replacements from original data and missing values (acts as dependent variable) are predicted using non-missing values (independent variable).

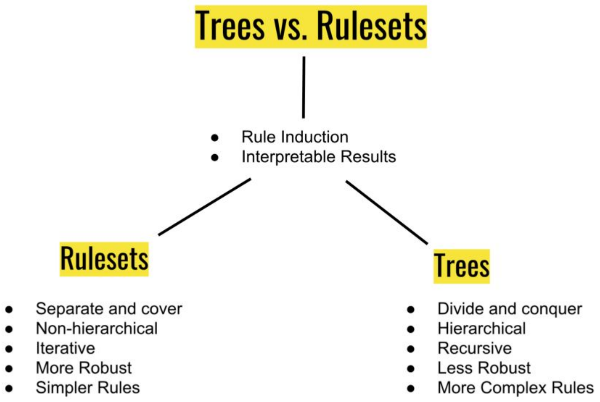
Then, it uses predictive mean matching (default) to impute missing values. Predictive mean matching works well for continuous and categorical (binary & multi-level) without the need for computing residuals and maximum likelihood fit.

Here are some important highlights of this package:

It assumes linearity in the variables being predicted.
Fisher's optimum scoring method is used for predicting categorical variables.

1. Initialize NAs to values from a random sample of the variable
2. For variables containing NAs, draw a sample with replacement from complete portion of the dataset (bootstrapping)
3. Fit a flexible additive model to predict the target variable while finding the optimum transformation of it.
4. Impute each missing value of target variable with observed variable (pmm)

- **Self Organizing Maps (SOMs)**
    - Heatmap grouping similar values of the same variables together
    - Dimension reduction because output is 2D map
    - https://www.superdatascience.com/blogs/the-ultimate-guide-to-self-organizing-maps-soms
    - SOM + K-means
        - https://www.shanelynn.ie/self-organising-maps-for-customer-segmentation-using-r/

- **C4.5**
    - Not imputation technique available in R
    - https://www.researchgate.net/publication/236635604_Data_Quality_Improvement_by_Imputation_of_Missing_Values
    - Rule sets (similar to decision trees)
    - Not hierarchical/ordered
    - 

Trees vs. Rulesets

- Rule Induction
- Interpretable Results

Rulesets

- Separate and cover
- Non-hierarchical
- Iterative
- More Robust
- Simpler Rules

Trees

- Divide and conquer
- Hierarchical
- Recursive
- Less Robust
- More Complex Rules

Pg 7: https://arxiv.org/pdf/1805.10572.pdf

Comparison methods

- Filter out complete observations (**this is ground truth**), split into test and train set for imputation modelling. Generate some missing values in the test set and calculate the accuracy of the model. Look at MAE between true and imputed value (lower values = better imputation). Can look at RMSE too
    - Pg 17: http://www.mit.edu/~dbertsim/papers/Machine%20Learning%20under%20a%20Modern%20Optimization%20Lens/From%20Predictive%20Methods%20to%20Missing%20Data%20Imputation.pdf

https://stefvanbuuren.name/fimd/sec-evaluation.html:

- Bias
    - Difference between expected value and actual value
- Coverage rate
    - Proportion of confidence intervals containing true value
- Average width
    - Average width of confidence interval is indicator of statistical efficiency. Length should be as small as possible, but not so small that CR will fall below nominal level
- RMSE
    - How spread out residuals (prediction errors) are

Pfc = proportion of false classification

- NRMSE

        - Equals 0 means imputation is perfect
        - Close to 1 means results similar to those obtained in mean imputation

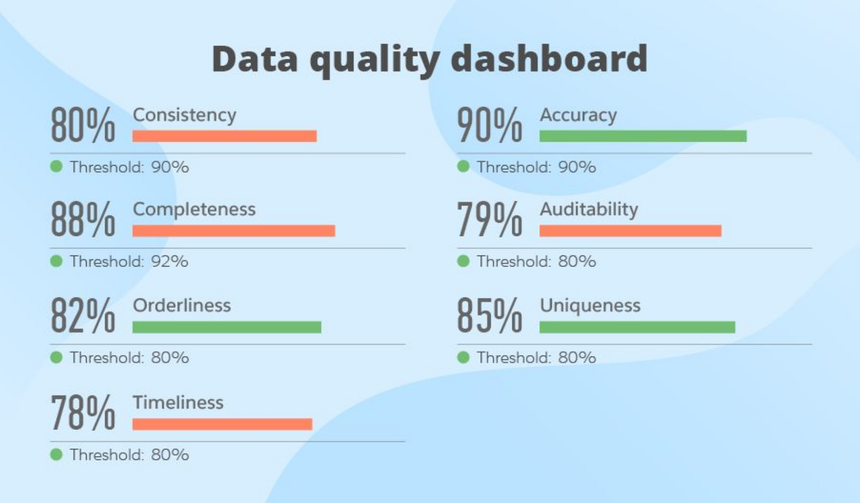
            - **Cox proportional hazards model**
            - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3939843/

- **Clustering with partial data to find similar units + measure of accuracy**
    - Constrained clustering: separate data into complete and incomplete sets
    - K-means with soft constraints
        - Soft constraint between 2 objects indicates their preference for or against their assignment to the same cluster
        - K-means seek to minimize total variance, while KSC minimizes combination of total variance and CV, a penalty for constraint violations
    - Expectation Maximization
        - Goal is to estimate mean and standard deviation of each cluster so as to maximize likelihood of observed data

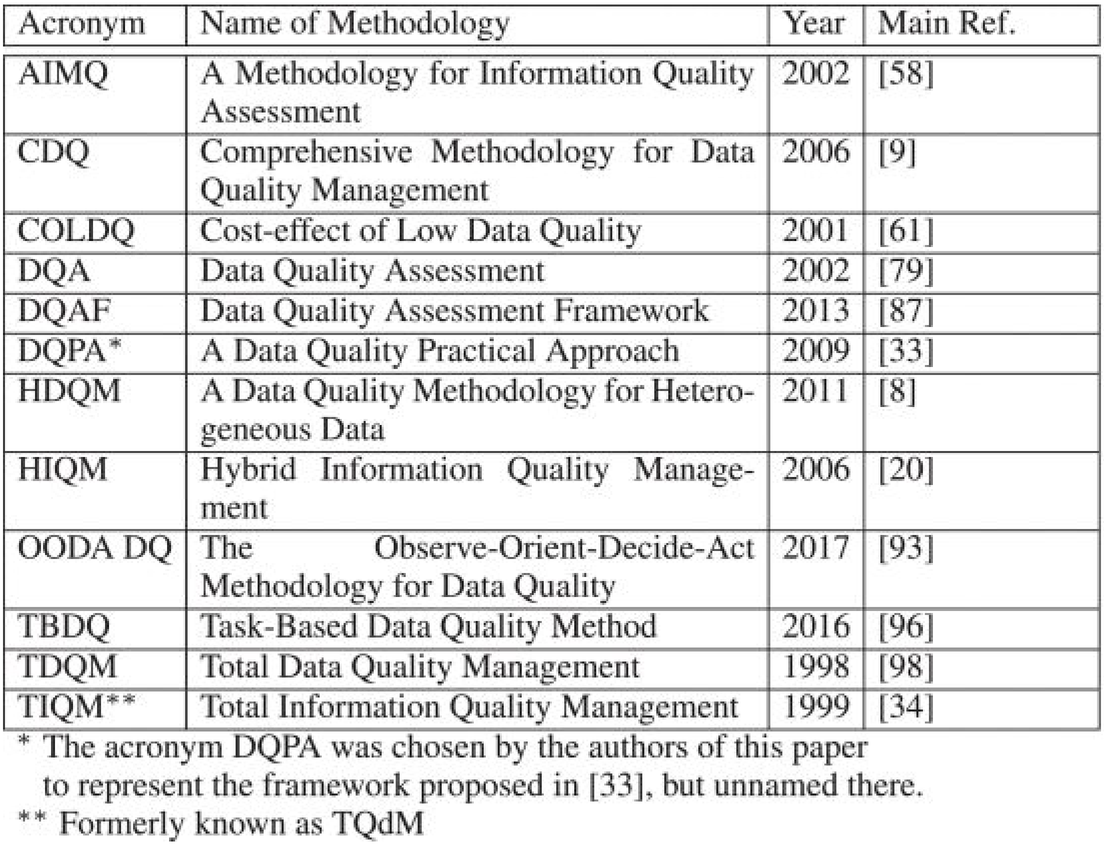
Imputation comes first followed by clustering???

Data Quality Model

- Ratio of data to errors
- Number of empty values
- Amount of dark data - data which is acquired through various computer network operations but not used in any manner to derive insights or for decision making. The ability of an organisation to collect data can exceed the throughput at which it can analyse the data.
- DAMA UK 6 dimensions of data quality
    - https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf
- Measuring timeliness
    - https://www.researchgate.net/publication/200047424_How_to_measure_data_quality_-_A_metric_based_approach
- Data rules



  -
  - What constitutes each measure? Definition?
- Weighted average
  - http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf

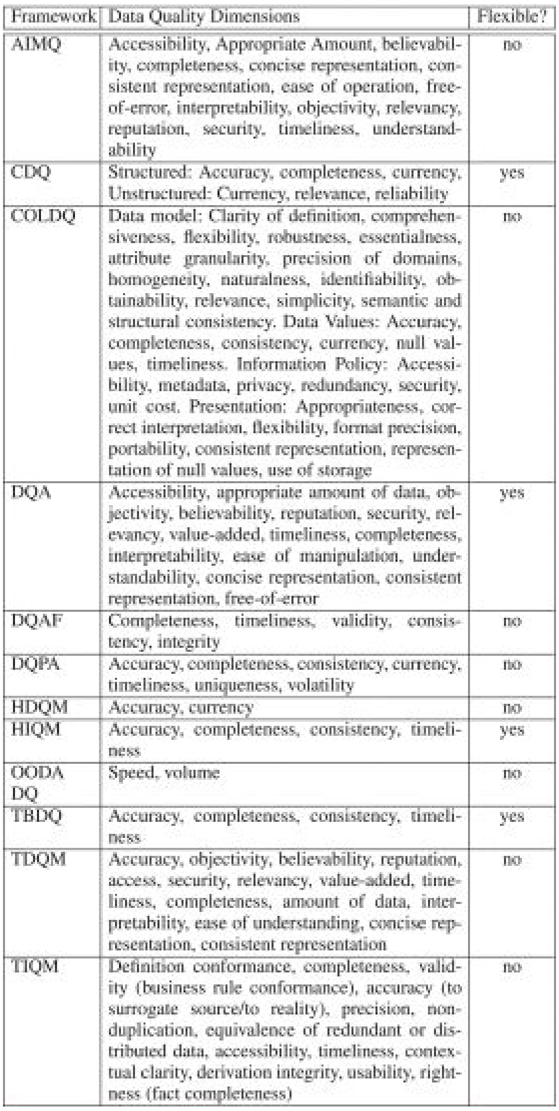


Acronym	Name of Methodology	Year	Main Ref.
AIMQ	A Methodology for Information Quality Assessment	2002	[58]
CDQ	Comprehensive Methodology for Data Quality Management	2006	[9]
COLDQ	Cost-effect of Low Data Quality	2001	[61]
DQA	Data Quality Assessment	2002	[79]
DQAF	Data Quality Assessment Framework	2013	[87]
DQPA*	A Data Quality Practical Approach	2009	[33]
HDQM	A Data Quality Methodology for Heterogeneous Data	2011	[8]
HIQM	Hybrid Information Quality Management	2006	[20]
OODA DQ	The Observe-Orient-Decide-Act Methodology for Data Quality	2017	[93]
TBDQ	Task-Based Data Quality Method	2016	[96]
TDQM	Total Data Quality Management	1998	[98]
TIQM**	Total Information Quality Management	1999	[34]

* The acronym DQPA was chosen by the authors of this paper to represent the framework proposed in [33], but unnamed there.

** Formerly known as TQdM

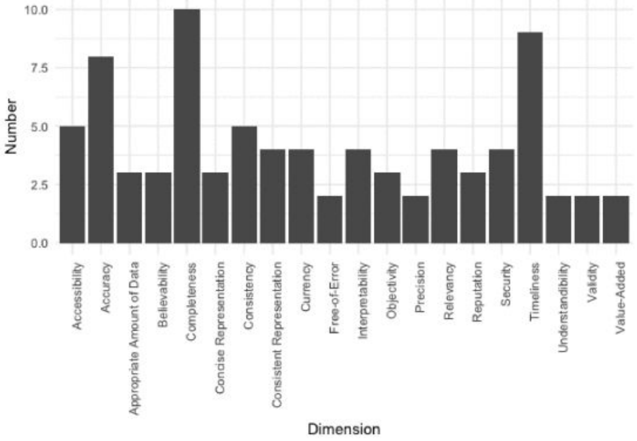
-
  - 12 types of DQ framework
  - 2019: https://ieeexplore.ieee.org/document/8642813



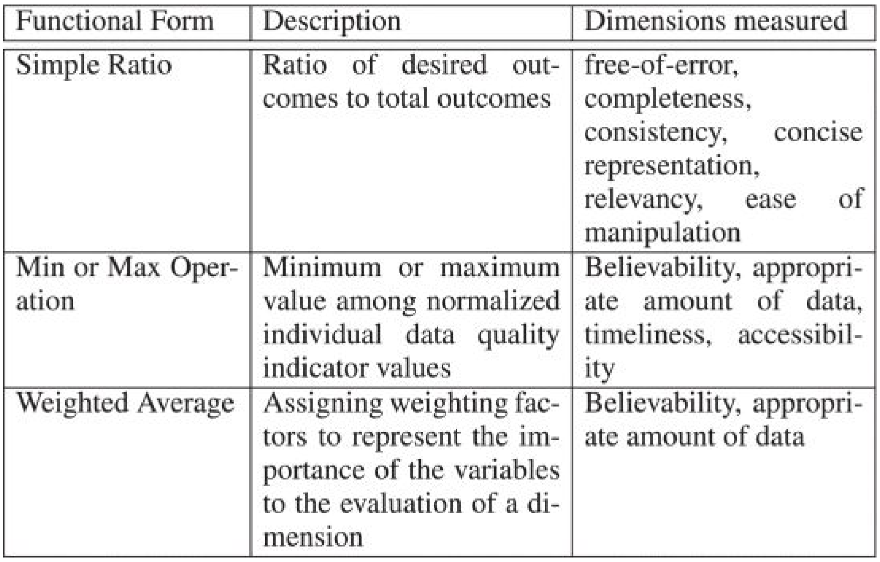
Framework	Data Quality Dimensions	Flexible?
AIMQ	Accessibility, Appropriate Amount, believability, completeness, concise representation, consistent representation, ease of operation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness, understandability	no
CDQ	Structured: Accuracy, completeness, currency, Unstructured: Currency, relevance, reliability	yes
COLDQ	Data model: Clarity of definition, comprehensiveness, flexibility, robustness, essentialness, attribute granularity, precision of domains, homogeneity, naturalness, identifiability, obtainability, relevance, simplicity, semantic and structural consistency. Data Values: Accuracy, completeness, consistency, currency, null values, timeliness. Information Policy: Accessibility, metadata, privacy, redundancy, security, unit cost. Presentation: Appropriateness, correct interpretation, flexibility, format precision, portability, consistent representation, representation of null values, use of storage	no
DQA	Accessibility, appropriate amount of data, objectivity, believability, reputation, security, relevancy, value-added, timeliness, completeness, interpretability, ease of manipulation, understandability, concise representation, consistent representation, free-of-error	yes
DQAF	Completeness, timeliness, validity, consistency, integrity	no
DQPA	Accuracy, completeness, consistency, currency, timeliness, uniqueness, volatility	no
HDQM	Accuracy, currency	no
HIQM	Accuracy, completeness, consistency, timeliness	yes
OODA DQ	Speed, volume	no
TBDQ	Accuracy, completeness, consistency, timeliness	yes
TDQM	Accuracy, objectivity, believability, reputation, access, security, relevancy, value-added, timeliness, completeness, amount of data, interpretability, ease of understanding, concise representation, consistent representation	no
TIQM	Definition conformance, completeness, validity (business rule conformance), accuracy (to surrogate source/to reality), precision, nonduplication, equivalence of redundant or distributed data, accessibility, timeliness, contextual clarity, derivation integrity, usability, rightness (fact completeness)	no

-

- CDQ, DQA, DQAF, DQPA, HIQM, TBDQ, TIQM

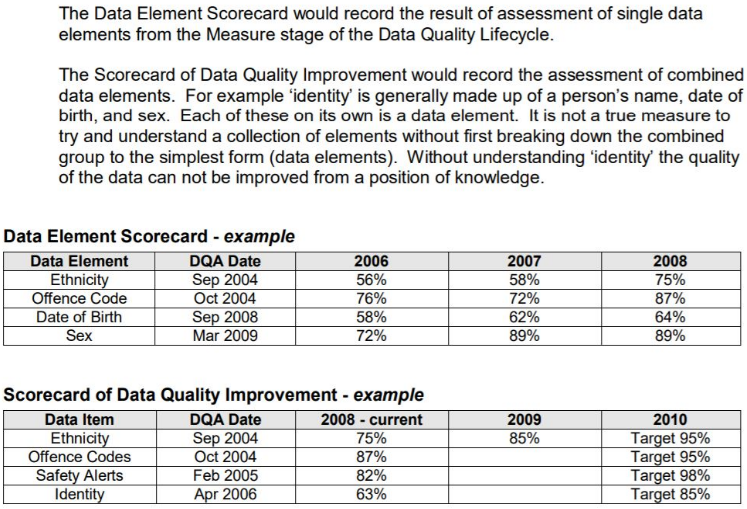


-



Functional Form	Description	Dimensions measured
Simple Ratio	Ratio of desired outcomes to total outcomes	free-of-error, completeness, consistency, concise representation, relevancy, ease of manipulation
Min or Max Operation	Minimum or maximum value among normalized individual data quality indicator values	Believability, appropriate amount of data, timeliness, accessibility
Weighted Average	Assigning weighting factors to represent the importance of the variables to the evaluation of a dimension	Believability, appropriate amount of data

    -
- Scorecards



The Data Element Scorecard would record the result of assessment of single data elements from the Measure stage of the Data Quality Lifecycle.

The Scorecard of Data Quality Improvement would record the assessment of combined data elements. For example 'identity' is generally made up of a person's name, date of birth, and sex. Each of these on its own is a data element. It is not a true measure to try and understand a collection of elements without first breaking down the combined group to the simplest form (data elements). Without understanding 'identity' the quality of the data can not be improved from a position of knowledge.

Data Element Scorecard - *example*

Data Element	DQA Date	2006	2007	2008
Ethnicity	Sep 2004	56%	58%	75%
Offence Code	Oct 2004	76%	72%	87%
Date of Birth	Sep 2008	58%	62%	64%
Sex	Mar 2009	72%	89%	89%

Scorecard of Data Quality Improvement - *example*

Data Item	DQA Date	2008 - current	2009	2010
Ethnicity	Sep 2004	75%	85%	Target 95%
Offence Codes	Oct 2004	87%		Target 95%
Safety Alerts	Feb 2005	82%		Target 98%
Identity	Apr 2006	63%		Target 85%

    -
    - https://static1.squarespace.com/static/5e21c300ec15d34ee6e45969/t/5e635de3b6faeb732cde6b11/1583570435364/New+Zealand+Ministry+of+Justice+Data+Quality+Framework+PDF

- UQ
    - Two branches: Inverse and forward problems
    - We look at forward problems: Y|X
    - https://www.frontiersin.org/articles/10.3389/fninf.2018.00049/full#F4
- Prior distribution
    - Distribution before you see data
- Posterior distribution
    - Distribution after you see data
- Similar use case:
    - https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2832-3
    - Dilema over whether to impute or cluster first. Use Monte Carlo?
- Monte carlo

- Stochastic collection with global/local polynomials
- Stochastic Galerkin
- Overview of imputation
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/
- Bayesian Statistics: One way is to compare modelling results before and after imputation
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4515885/
    - Main purpose of MICE as well
    - Posterior predictive checking: delete random values from imputed data and re-impute the data, and compare with the ones prior to deletion
    - https://www.researchgate.net/publication/327968815_Improving_performance_of_classification_on_incomplete_data_using_feature_selection_and_clustering
        - Feature selection, data imputation then clustering
- Root mean square error (RMSE), mean relative error (MRE), root mean square deviation (RMSD), and mean relative deviation (MRD)
    - file:///C:/Users/d-ebb/Downloads/applsci-09-00204.pdf
    - **Maybe can compare 60% clustering results with imputed dataset cluster results?**
    - https://www.hindawi.com/journals/wcmc/2019/4039758/

- Capture reasons for MNAR/MAR before sensitivity analysis
    - https://www.ncbi.nlm.nih.gov/books/NBK209900/

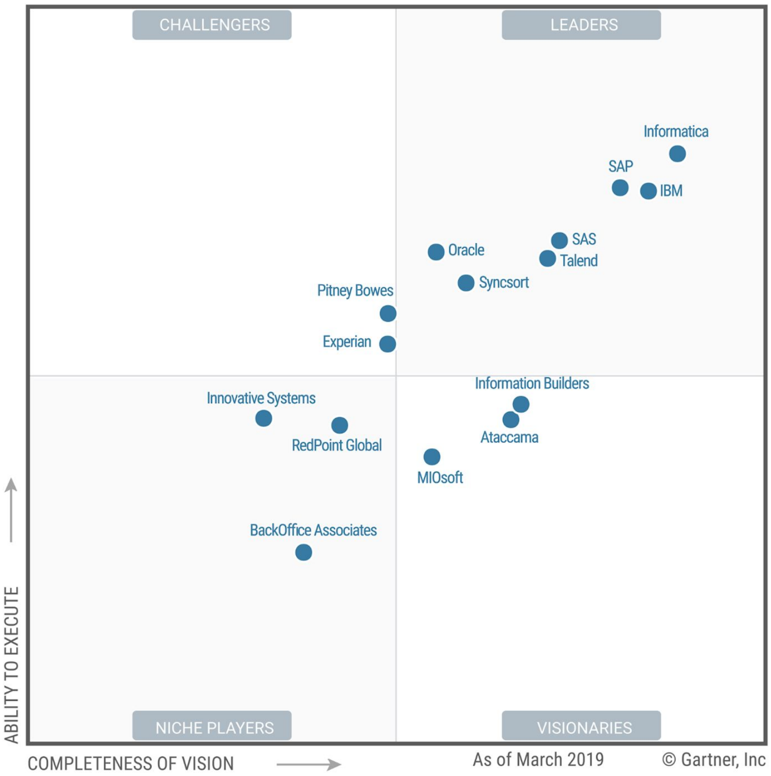
- Karhunen-Loeve expansion
- Monte Carlo Sampling

DQ Frameworks

Framing the topic:

https://searchdatamanagement.techtarget.com/definition/data-quality

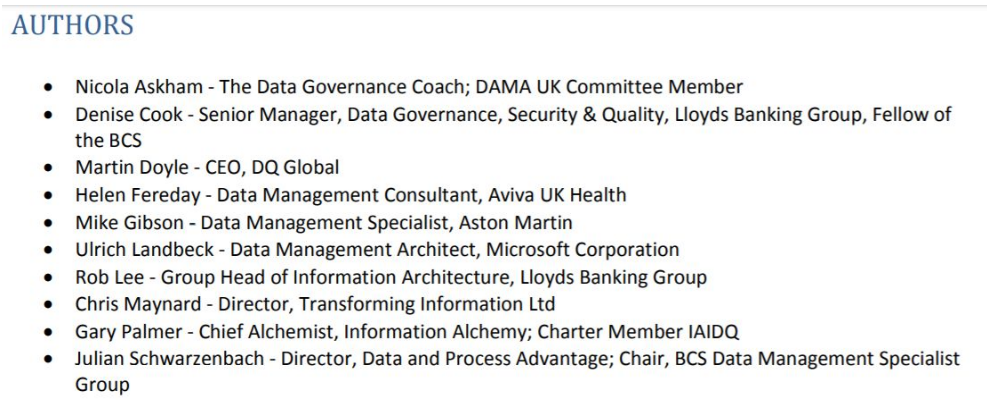
Figure 1. Magic Quadrant for Data Quality Tools



Source: Gartner (March 2019)

-
- IBM: https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ia.product.doc/topics/c_quality_score.html

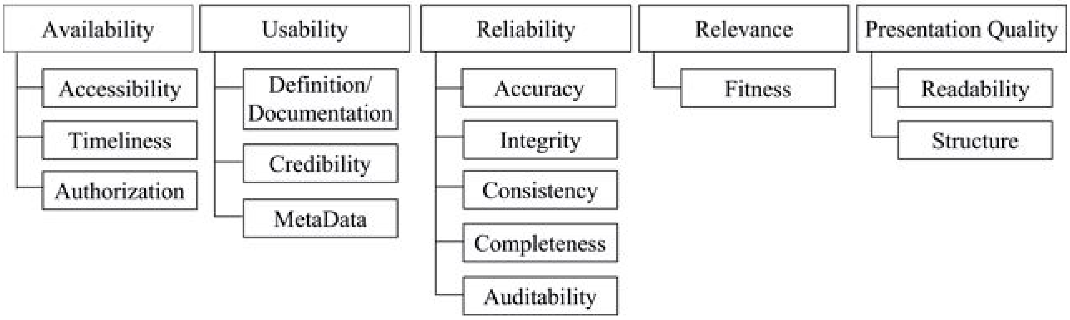
- DAMA UK 6 dimensions
    - https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf
    - Authored by a few big names

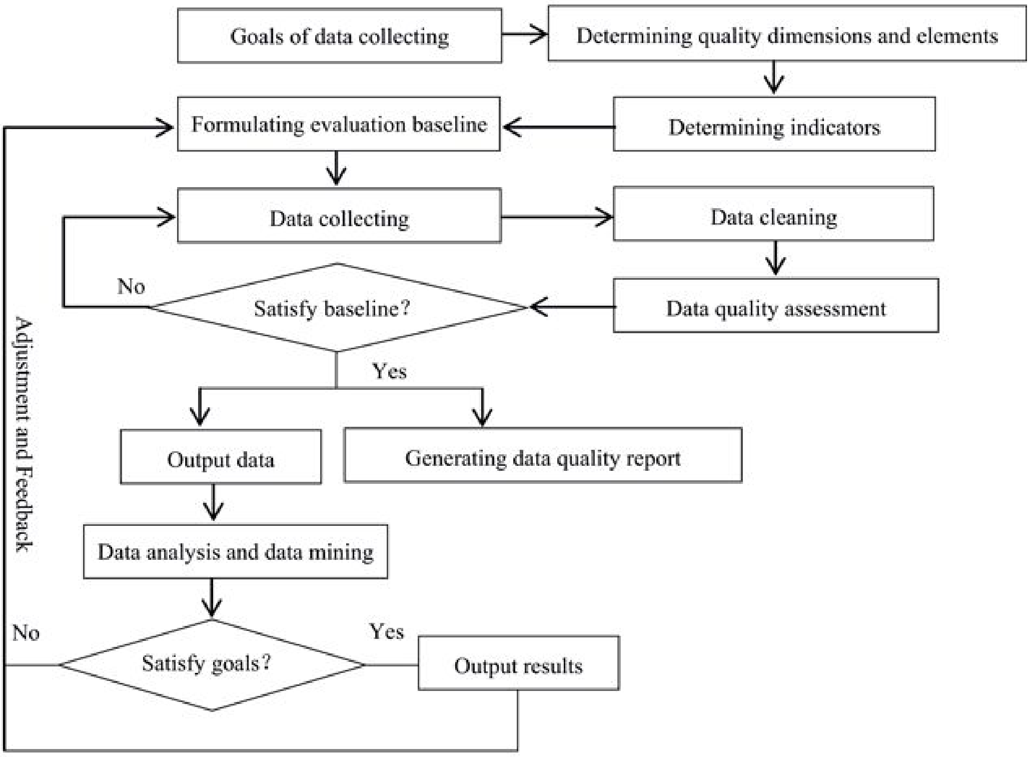


AUTHORS

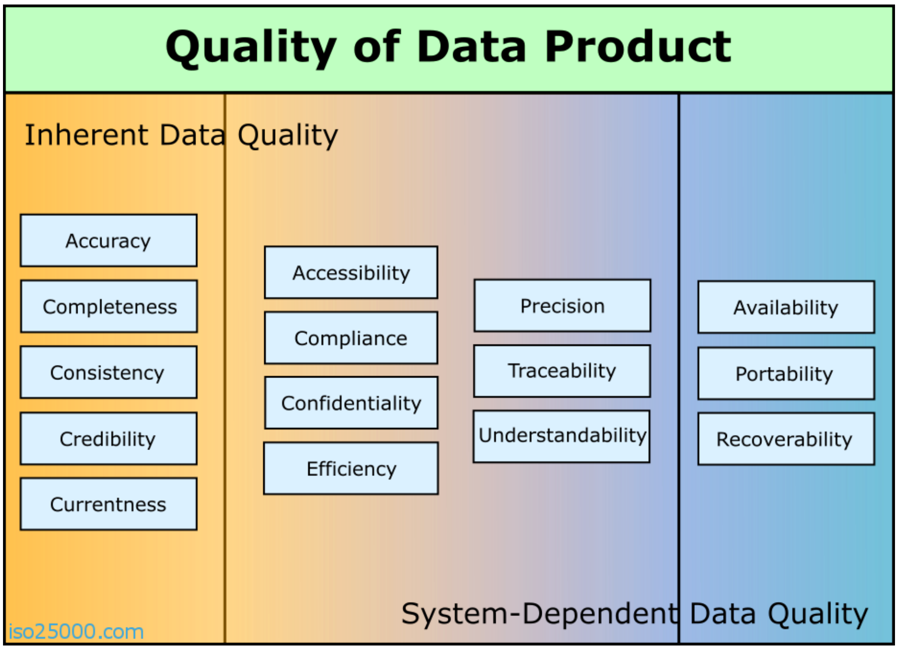
- Nicola Askham - The Data Governance Coach; DAMA UK Committee Member
- Denise Cook - Senior Manager, Data Governance, Security & Quality, Lloyds Banking Group, Fellow of the BCS
- Martin Doyle - CEO, DQ Global
- Helen Fereday - Data Management Consultant, Aviva UK Health
- Mike Gibson - Data Management Specialist, Aston Martin
- Ulrich Landbeck - Data Management Architect, Microsoft Corporation
- Rob Lee - Group Head of Information Architecture, Lloyds Banking Group
- Chris Maynard - Director, Transforming Information Ltd
- Gary Palmer - Chief Alchemist, Information Alchemy; Charter Member IAIDQ
- Julian Schwarzenbach - Director, Data and Process Advantage; Chair, BCS Data Management Specialist Group

-

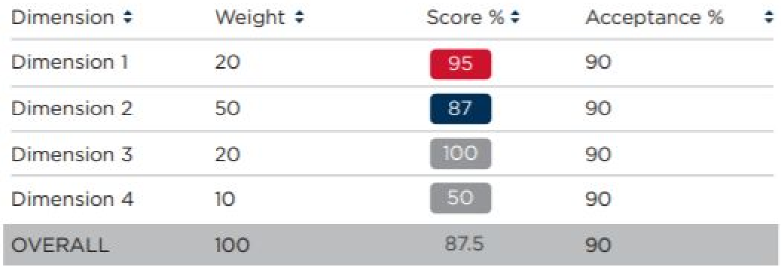
        - Closely similar case at M&S: https://it.ojp.gov/documents/Informatica_Whitepaper_Monitoring_DQ_Using_Metrics.pdf
        - PwC framework: https://www.pwc.com/us/en/industries/financial-services/research-institute/blog/automation-data-quality-testing.html
    - 2013 (7 years of release)
- 

    - https://datascience.codata.org/articles/10.5334/dsj-2015-002/#B6
    - Framework:
    - 

- http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf
    - Functional forms
        - Simple ratio
        - min/max
        - Weighted average
- WHO's framework **[scorecards and actions to take]**
    - https://apps.who.int/iris/bitstream/handle/10665/259224/9789241512725-eng.pdf;jsessionid=4A8A527E48E8482500AB487B6FC2D5E7?sequence=1

    - Completeness & timeliness
    - Internal consistency
    - External consistency (agreement with other sources of data)
- ISO 250000 framework
    - 

    - https://iso25000.com/index.php/en/iso-25000-standards/iso-25012



Dimension ⇕	Weight ⇕	Score % ⇕	Acceptance % ⇕
Dimension 1	20	95	90
Dimension 2	50	87	90
Dimension 3	20	100	90
Dimension 4	10	50	90
OVERALL	100	87.5	90

Figure 5: Data Quality Dimension Scores and the Overall Data Quality Score (Weighted Average of Dimensional Scores)

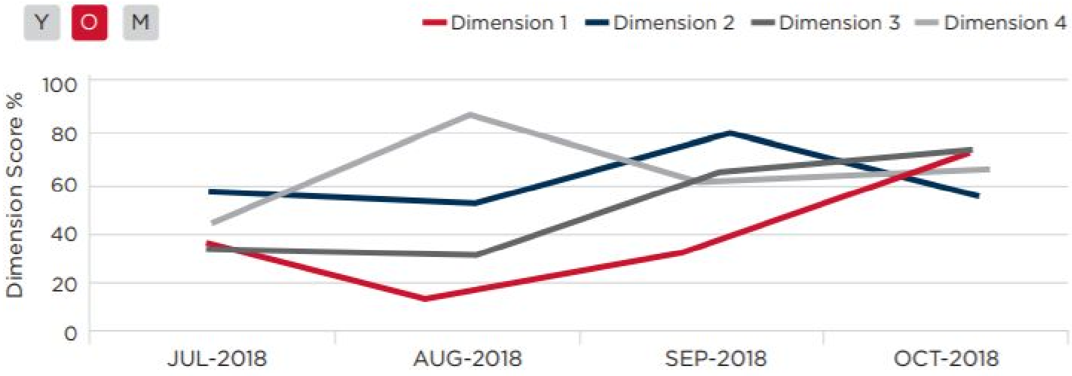
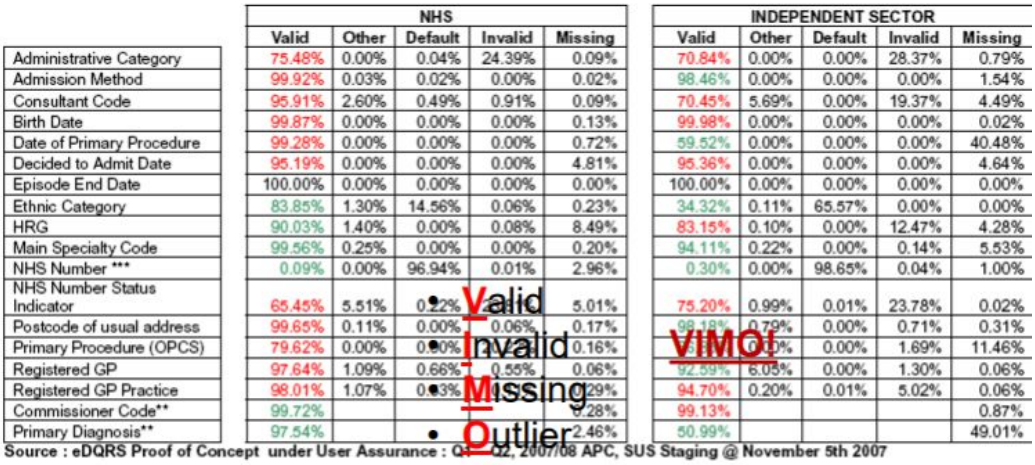


Figure 6: Data Quality Dimension Score Trends

    -

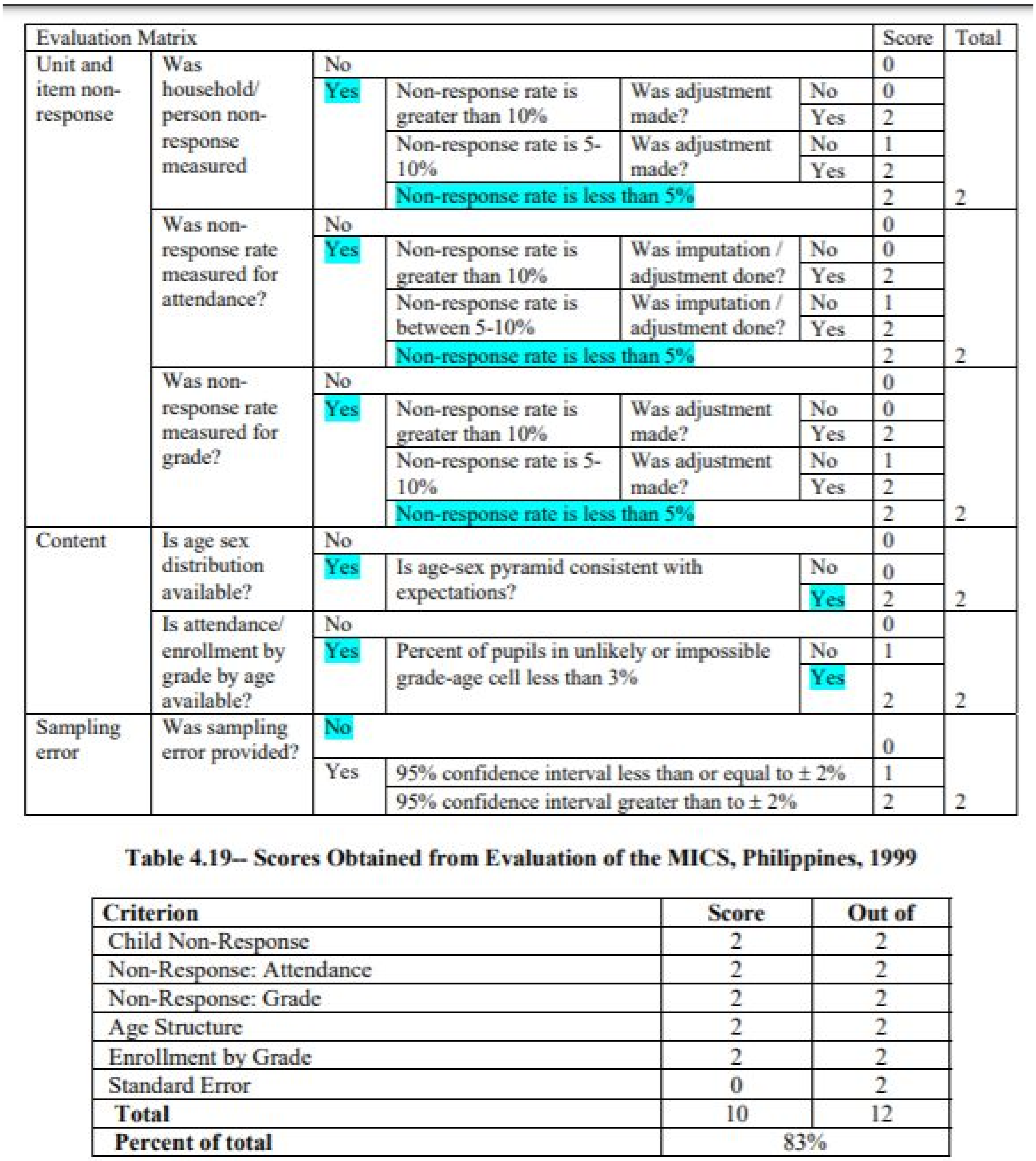
    - (2019) ISO 8000: https://ww2.eagle.org/content/dam/eagle/advisories-and-debriefs/data-quality-for-marine-offshore-application_19325.pdf
- Tabular form to showcase difference data dimensions



	NHS					INDEPENDENT SECTOR				
	Valid	Other	Default	Invalid	Missing	Valid	Other	Default	Invalid	Missing
Administrative Category	75.48%	0.00%	0.04%	24.39%	0.09%	70.84%	0.00%	0.00%	28.37%	0.79%
Admission Method	99.92%	0.03%	0.02%	0.00%	0.02%	98.46%	0.00%	0.00%	0.00%	1.54%
Consultant Code	95.91%	2.60%	0.49%	0.91%	0.09%	70.45%	5.69%	0.00%	19.37%	4.49%
Birth Date	99.87%	0.00%	0.00%	0.00%	0.13%	99.98%	0.00%	0.00%	0.00%	0.02%
Date of Primary Procedure	99.28%	0.00%	0.00%	0.00%	0.72%	59.52%	0.00%	0.00%	0.00%	40.48%
Decided to Admit Date	95.19%	0.00%	0.00%	0.00%	4.81%	95.36%	0.00%	0.00%	0.00%	4.64%
Episode End Date	100.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
Ethnic Category	83.85%	1.30%	14.56%	0.06%	0.23%	34.32%	0.11%	65.57%	0.00%	0.00%
HRG	90.03%	1.40%	0.00%	0.08%	8.49%	83.15%	0.10%	0.00%	12.47%	4.28%
Main Specialty Code	99.56%	0.25%	0.00%	0.00%	0.20%	94.11%	0.22%	0.00%	0.14%	5.53%
NHS Number ***	0.09%	0.00%	96.94%	0.01%	2.96%	0.30%	0.00%	98.65%	0.04%	1.00%
NHS Number Status Indicator	65.45%	5.51%	0.22%	[illegible]	5.01%	75.20%	0.99%	0.01%	23.78%	0.02%
Postcode of usual address	99.65%	0.11%	0.00%	0.06%	0.17%	98.18%	0.79%	0.00%	0.71%	0.31%
Primary Procedure (OPCS)	79.62%	0.00%	[illegible]	[illegible]	0.16%	[illegible]	[illegible]	0.00%	1.69%	11.46%
Registered GP	97.64%	1.09%	0.66%	0.55%	0.06%	92.59%	6.05%	0.00%	1.30%	0.06%
Registered GP Practice	98.01%	1.07%	[illegible]	[illegible]	[illegible]	94.70%	0.20%	0.01%	5.02%	0.06%
Commissioner Code**	99.72%				0.28%	99.13%				0.87%
Primary Diagnosis**	97.54%				2.46%	50.99%				49.01%

Source : eDQRS Proof of Concept under User Assurance : Q1 – Q2, 2007/08 APC, SUS Staging @ November 5th 2007

    -
    - https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Toronto-Data-Mining-Forum/Azimaee-DQAssurance-2012.pdf
- Among DFs propose to follow these 5 requirements for standardization in collection and measurement of metrics
    - https://epub.uni-regensburg.de/36889/1/Requirements%20for%20Data%20Quality%20Metrics.pdf



Evaluation Matrix						Score	Total
Unit and item non-response	Was household/ person non-response measured	No				0	
		Yes	Non-response rate is greater than 10%	Was adjustment made?	No	0	
					Yes	2	
			Non-response rate is 5-10%	Was adjustment made?	No	1	
					Yes	2	
			Non-response rate is less than 5%			2	2
	Was non-response rate measured for attendance?	No				0	
		Yes	Non-response rate is greater than 10%	Was imputation / adjustment done?	No	0	
					Yes	2	
			Non-response rate is between 5-10%	Was imputation / adjustment done?	No	1	
					Yes	2	
			Non-response rate is less than 5%			2	2
	Was non-response rate measured for grade?	No				0	
		Yes	Non-response rate is greater than 10%	Was adjustment made?	No	0	
					Yes	2	
			Non-response rate is 5-10%	Was adjustment made?	No	1	
					Yes	2	
			Non-response rate is less than 5%			2	2
Content	Is age sex distribution available?	No				0	
		Yes	Is age-sex pyramid consistent with expectations?		No	0	
					Yes	2	2
	Is attendance/ enrollment by grade by age available?	No				0	
		Yes	Percent of pupils in unlikely or impossible grade-age cell less than 3%		No	1	
					Yes	2	2
Sampling error	Was sampling error provided?	No				0	
		Yes	95% confidence interval less than or equal to ± 2%			1	
			95% confidence interval greater than to ± 2%			2	2

Table 4.19-- Scores Obtained from Evaluation of the MICS, Philippines, 1999

Criterion	Score	Out of
Child Non-Response	2	2
Non-Response: Attendance	2	2
Non-Response: Grade	2	2
Age Structure	2	2
Enrollment by Grade	2	2
Standard Error	0	2
Total	10	12
Percent of total	83%	

-
    - https://www.epdc.org/sites/default/files/documents/Methodology_for_Evaluating_Data_Quality.pdf

Clustering similar observations without labels

Aim is to find similar units to those that are registered as having caught for being an illegal dorm.

No labels.

3 ways:

1. Compute distance matrix

    a. Find closest observations for every observation
    b. But computationally expensive
2. Impute and do clustering
3. Don’t impute and do clustering

Choose 1...

https://hal.archives-ouvertes.fr/hal-01355189v2/document
- Do k-means for labelled data
- Use the centroids as centroids for data with all unlabelled + labelled data

Determine number of clusters
https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc
- Empirical method
    - Sq root of N/2
- Elbow method

Exclusive clustering
- SVM
- K-medoids
    - Similar to k-means but instead of cluster mean being used, an actual data point is used as centre of cluster
    - Different from kmeans
        - Accepts dissimilarity matrix
            - **BUT** takes in a lot of memory. O(n^2) https://datascience.stackexchange.com/questions/22/k-means-clustering-for-mixed-numeric-and-categorical-data
        - More robust because it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances
        - Provides novel graphical display
        - Allows to select number of clusters
- KNN
    - Cons https://www.math.leidenuniv.nl/scripties/MasterWilson.pdf
        - Specify number of clusters
        - Initial assignment of observations can affect outcome of clusters
        - Clusters assumed to be spherically shaped
        - Should not be used because it requires for continuous variables. Even if encoded it doesn’t make sense in distance computation e.g. NY number 3, LA number 8, distance is 5 but it has nothing to do with difference between the 2 values
        - **Can hot encode followed by clustering. But will issue of curse of dimensionality**

Overlapping clustering
- Fuzzy c-means (model based clustering)
    - Specifies uncertainty in clustering/classification

        - Each point may belong to >1 cluster with different degrees of membership

Hierarchical clustering

- Hierarchical clustering
    - Cons
        - One way algorithm. Observation cannot be reassigned once partition or division occurred
        - Computationally expensive
        - Can only be done on continuous data

Probabilistic clustering

- Bayesian HC
    - Instead of physical distance, probabilistic distance between data points is used
- EM
    - Assumes multivariate normal distributions (strong assumption so no go)

==

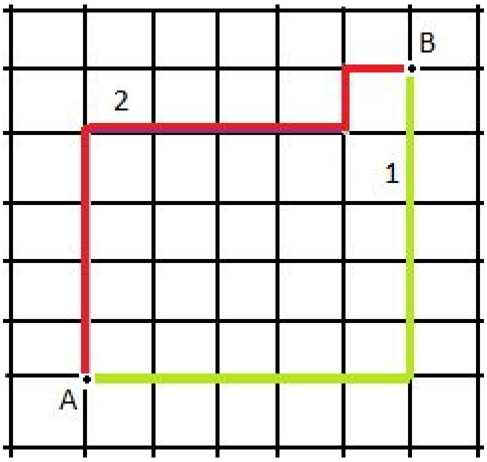
- GLCM/GLRLM for feature extraction
- k-POD
- For multiple imputed datasets
    - Run clustering and each assign modal cluster to observations
- KSC http://www.litech.org/~wkiri/Papers/wagstaff-missing-ifcs04.pdf
    - Clustering soft constraints: indicate how strongly a pair of items should or should not be grouped together (approximate information)

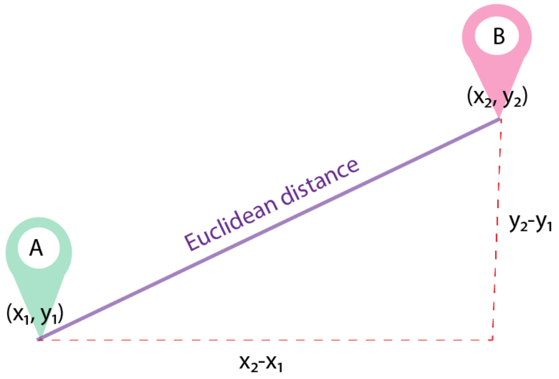
Evaluate outputs of clustering

- With ground truth (extrinsic measures)
    - ARI: adjusted rand index
    - Fowlkes-Mallows scores
    - Mutual information based scores
    - Homogeneity
    - Completeness
    - V-measure
- Without ground truth (intrinsic measures)
    - Silhouette Coefficient
        - Calculate how similar each observations is with the cluster relative to other clusters
        - Estimates average distance between clusters
        - -1 to 1. Close to 1 = observation is well matched to assigned cluster, close to 0 means borderline between 2 clusters. Close to -1 means observations may be assigned to wrong cluster
        - Should be > 0.4 https://stats.stackexchange.com/questions/320831/interpreting-silhouette-plot-for-cluster-analysis

    - Dunn Index
        - Should be maximized
        - Higher means that clusters are compact and well-separated from other clusters
        - Minimum inter-cluster distance divided by maximum cluster size
        - If the data set contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large. Thus, Dunn index should be maximized.
        - Ie. larger inter-cluster distance and smaller cluster size will lead to higher DI value
    - Daives-Bouldin Index
        - Cannot use. Cos it is only for Manhattan/Euclidean distance

Distance Metrics

- Minkowski Distance
    - Generalized distance metric
- Manhattan Distance
    - City block distance/taxicab geometry
    - Calculate distance between two data points in a grid-like path
    - 

    - Usually used with high dimensionality
- Euclidean Distance
    - Straight line distance between two data points in a plane
    - 

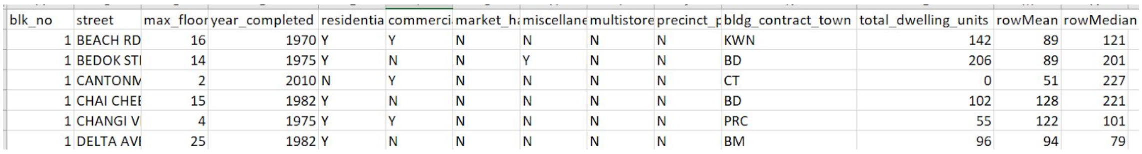
    - Only for continuous data https://www.r-bloggers.com/clustering-mixed-data-types-in-r/
- Hamming Distance
    - To compare two binary data strings
    - Perform XOR operation and count total number of 1s in the string
- Cosine distance & Cosine similarity
    - Find similarities between two data points
    - Inverse relationship, points closer to each other have lower Cosine similarity score
- Gower Distance
    - For each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1
    - Quantitative (intervale): range-normalized Manhattan distance
    - Ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
    - Nominal: variables of k are first converted into k binary columns and then Dice coefficient is used

- Single linkage
    - Shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than two observations within the clusters
- Complete linkage
    - Distance between farthest pair of observations in two clusters
- Average linkage
    - Distance between each pair of observations in each cluster are added up and divided by number of pairs
- Centroid linkage
    - Distance between centroids of two clusters

<u>Questions to answer (in order of priority)</u>

1. DQ Framework
2. Imputation techniques for partial data (resi is 60% complete, hopefully its possible to accurately impute the other 40% incomplete rows) & how imputation accuracy quantification
3. Clustering to identify similar units - question here is to use complete data or complete+imputed data?

Week 1:

- Data quality frameworks & DQI
    - Compare the different frameworks & recommend
    - Currently in the DQ dimension completeness & accuracy are most relevant
- Uncertainty Quantification
    - See if its applicable to our use case: mainly to solve imputation problems now
- Came up with dummy data similar to URA Resi's
    - Used data from https://data.gov.sg/dataset/hdb-property-information
    - 

blk_no	street	max_floor	year_completed	residentia	commerci	market_h	miscellane	multistore	precinct_p	bldg_contract_town	total_dwelling_units	rowMean	rowMedian
1	BEACH RD	16	1970	Y	Y	N	N	N	N	KWN	142	89	121
1	BEDOK STI	14	1975	Y	N	N	Y	N	N	BD	206	89	201
1	CANTONM	2	2010	N	Y	N	N	N	N	CT	0	51	227
1	CHAI CHEE	15	1982	Y	N	N	N	N	N	BD	102	128	221
1	CHANGI V	4	1975	Y	Y	N	N	N	N	PRC	55	122	101
1	DELTA AVI	25	1982	Y	N	N	N	N	N	BM	96	94	79

    - rowMedian rowMean represents the PUB billing data for now

Week 2:

- Continue work on UQ (because now it looks more promising than DQ framework since its more scientific. But still researching and trying to understand it)

Week 3:

- Main focus is still DQ frameworks & quantifying + looking at qn 2
- Imputation (follow up from the work I did previous weeks which is researching and playing around with different imputation techniques) to be applied on simulated data and compare results
    - C4.5
    - Naive Bayes
    - KNN (iterative, sequential, regular)
    - Bayesian PCA
    - SVD
    - SVM
    - SOM
- One of the more interesting imputation techniques found: http://www.mit.edu/~dbertsim/papers/Machine%20Learning%20under%20a%20Modern%20Optimization%20Lens/From%20Predictive%20Methods%20to%20Missing%20Data%20Imputation.pdf
- UQ to be applied on imputation models hopefully(?)

Week 4:

- Main focus is still DQ frameworks & quantifying + looking at qn 2 & 3